Review of Educational Research

VOL. XXIX, NO. 1 FEBRUARY 1959

EDUCATIONAL AND PSYCHOLOGICAL TESTING

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

A Department of the
NATIONAL EDUCATION ASSOCIATION OF THE UNITED STATES
1201 Sixteenth St., N.W., Washington 6, D. C.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

THIS ASSOCIATION is composed of persons engaged in technical research in education, including directors of research in school systems, instructors in educational institutions, and research workers connected with private educational agencies.

Applications for membership should be sent to the secretary-treasurer. Upon approval by a committee of the Association, persons applying will be invited to become members.

Subscriptions to the REVIEW should be sent to the secretary-treasurer (note address above).

Orders for one or more publications, accompanied by funds in payment, should be sent to the American Educational Research Association, 1201 Sixteenth St., N. W., Washington 6, D. C. For a list of topics see the back inside cover page.

Active and associate members of the Association pay dues of $10 annually. Of this amount $7 is for subscription to the REVIEW. The REVIEW is published in February, April, June, October, and December. Beginning with the October 1957 issue, single copies are priced at $2.

Entered as second-class matter, April 10, 1931, at the post office at Washington, D. C., under the Act of August 24, 1912.

REVIEW OF EDUCATIONAL RESEARCH

Official Publication of the American Educational Research Association. Contents are listed in the Education Index.

The Library of Congress catalogue entry for this publication appears on page 2.

Vol. XXIX, No. 1 **February 1959**

Educational and Psychological Testing

Reviews the literature for the three-year period since the issuance of Vol. XXVI, No. 1, February 1956.

TABLE OF CONTENTS

Review of educational research. v. 1–
Jan. 1931–
Washington, American Educational Research Association.

v. 24 cm. 5 no. a year.

Each number is devoted to a specific educational subject, and includes bibliographies.
INDEXES:
Vols. 1–12, 1931–42. 1 v. (Special issue, Dec. 1944)

1. Education—Period. 2. Education—U. S. 3. Education—Bibl. I. American Educational Research Association.

L11.R35 370.5 33–19994 rev 2

Library of Congress

This issue of the REVIEW was prepared by the Committee on Educational and Psychological Testing.

WILLIAM B. MICHAEL, *Chairman*, University of Southern California, Los Angeles, California
DOROTHY MANLEY COLLETT, La Verne College, La Verne, California
ROBERT L. EBEL, Educational Testing Service, Princeton, New Jersey
BENJAMIN FRUCHTER, The University of Texas, Austin, Texas
DAVID R. KRATHWOHL, Michigan State University, East Lansing, Michigan
SAMUEL T. MAYO, Loyola University, Chicago, Illinois
JOHN W. M. ROTHNEY, University of Wisconsin, Madison, Wisconsin
WILLARD G. WARRINGTON, Michigan State University, East Lansing, Michigan

with the assistance of

CHERRY ANN CLARK, The Meyers Clinic, Los Angeles, California, and the Claremont Graduate School, Claremont, California
WILLIAM COLEMAN, Systems Development Corporation, Santa Monica, California
J. P. GUILFORD, University of Southern California, Los Angeles, California
ROBERT A. HEIMANN, Arizona State College, Tempe, Arizona
ROBERT E. HILL, JR., Ball State Teachers College, Muncie, Indiana
H. PAUL KELLEY, The University of Texas, Austin, Texas
JOE L. SAUPE, Michigan State University, East Lansing, Michigan

EDUCATIONAL PRESS ASSOCIATION OF AMERICA

INTRODUCTION

ALTHOUGH following the pattern of organization of the most recent issue of the REVIEW concerned with problems of educational and psychological testing (Volume XXVI, No. 1), the current issue contains two additional chapters: one upon tests of general mental ability and another upon the challenging area of construct validity. No attempt will be made in this introduction either to summarize the contents of the chapters or to present trends as the reader will find in the chapters themselves evaluative material as well as descriptive coverage of pertinent research literature.

During the preparation of the current issue of the REVIEW the chairman followed the practice of his predecessor, Max Englehart, of writing to more than 200 members of the American Educational Research Association in order to request copies of references to materials such as bulletins or monographs upon educational and psychological testing that are not readily found in the more familiar sources. The Chairman of the Committee evaluated the various references and publications received and sent to respective chapter authors those items that he thought relevant. Appreciation is expressed by the Committee and its helpers to the many AERA members who gave so generously of their time and efforts.

Because of tremendous growth in the number of articles published in professional journals it has been necessary to be highly selective. It is hoped that a minimum number of significant contributions have been overlooked or omitted from each of the eight chapters.

WILLIAM B. MICHAEL, *Chairman*
Committee on Educational and Psychological Testing

CHAPTER I

Testing and the Use of Test Results

SAMUEL T. MAYO

THIS introductory chapter presents a general overview of testing. Three aspects are treated: developments contributing to the improvement of tests and testing, developments in testing programs, and sources of information on testing.*

The last three years saw several reviews of testing history as well as critical evaluations of the philosophy, theory, and practice of testing; illustrations are the contributions of Cronbach and Gleser (12), Kavruck (43), Traxler (77), and Wrightstone and others (85). Cronbach and Gleser asked whether traditional psychometric theory is perhaps outmoded. After several years of trying to devise a testing model based upon information theory, Cronbach and his associates turned to decision theory as being more adequate. They touched upon such basic issues in the improvement of testing as optimum selection ratio, optimum length of a single test, optimum size of a test battery, sequential testing, and the bandwidth-fidelity dilemma.

In an invited address before the annual meeting of the American Psychological Association in 1957, Professor Philip N. Vernon of the University of London observed that despite 25 years of the wide use of tests in education and in the military services and in spite of considerable gains in test theory, the practical efficiency of testing was still disappointing. He further pointed out that current tests involved many sources of variance other than the constructs at which they were aimed. He suggested that after more thorough exploration of components, a relatively short list of ability and personality factors could be devised which would cover much of the variance in performance criteria used in making practical decisions.

There seemed to be a number of counter-trends toward correcting the previous neglect of test validity. More attention to construct validity was given in the newer test manuals. An entire chapter of the present issue is devoted to construct validity.

It was encouraging to note evidence of careful attention to adequate criterion variables in a number of references such as those by Flanagan (32); Macaluso and Dailey (51); Perloff (63); Stein (74); Stuit, Helmstadter, and Frederiksen (75); and Wilson (83).

There was widespread evidence of increased efforts to educate test users in the better understanding of the purposes, characteristics, and interpre-

*A supplemental bibliography may be obtained free from the author while his supply lasts.

tation of tests. This was manifested not only in recommendations for increased course work in testing for prospective teachers, but also in augmented inservice education of all test users. In spite of progress in these directions, it was implied that not nearly enough was being done to educate prospective teachers in measurement.

Noll (60) surveyed requirements for measurement courses for certification in the various states and the course work offered in measurement in 80 selected teacher-training institutions of four types: large public, large private, state teachers colleges, and liberal arts colleges. He found that 83 percent offered an introductory course in measurement. Of these, however, only about 14 percent required the course of all undergraduates preparing to teach; up to 21 percent required it of undergraduates preparing for certain types of certificates. Only about 10 percent of the states specified a course in measurement for certification, and it was even rare that states recommended such a course as an elective.

Under the auspices of the Committee on Test Utilization of the National Council on Measurements Used in Education, Allen (1) surveyed measurement course offerings and opinions relative thereto in 288 teacher-training institutions, obtaining results similar to Noll's. She found also that a majority of the institutions had reference libraries of standardized tests and reported adequate assistance from test publishers. There was less consensus as to the adequacy of instructional materials and methods, and some specific suggestions for improving these were cited from questionnaire responses.

Diederich (15, 17, 18), reported a practicum in item analysis in his introductory measurement classes. Several other authors who gave attention to practical suggestions for exercise writing and simplified item analysis for teachers included Engelhart (30), Nedelsky (58), Schumacher (67), and Stecklein (73). A number of testing and research bureaus of large universities and colleges circulated form letters to acquaint their faculties with their consulting services on course examinations and also distributed bulletins designed to acquaint the faculties with principles of test construction. Among these were the Chicago City Junior College, Michigan State University, the University of Minnesota, and the University of Southern California. Other authors discussed steps that could be taken to make test results more meaningful and more useful to teachers, administrators, and students. Among these were Allison and Helmstadter (2), Coleman (11), Diederich (16), Doppelt (20, 21), Gustad (39), Hart (40), Seashore (70), and Wesman (81, 82).

More attention was given to psychological factors in testing. Rimoldi (65, 66) described a new type of problem-solving item form, which emphasizes the *processes* of learning rather than *products*, and he also explored several new ways of scoring responses. In a handbook for college teachers Dressel and Hill (23) related "critical thinking" as one of the important objectives of education to problems of instruction and evaluation. Dressel

(22) also considered a large number of psychological factors in a series of institutional research studies, for example, biases in academic marking, students' fear of "blowing up" on examinations, and immediate knowledge of results in answering test items.

A number of developments were made in evaluative instruments and techniques other than tests. Flanagan, Pumroy, and Tuska (35) applied the critical incident technique to the development of a personality criterion instrument, and the same instrument was offered as an aid to behavioral record keeping in the elementary grades (33, 34). North (61) studied a biographical inventory experimentally in four metropolitan high schools. A new method of scaling categorical data, such as personal history blanks, profile scores, and inventory items, was proposed by du Mas (24).

It was encouraging to note that the amount of funds available for research on testing and evaluation instruments greatly increased. There were too many sources to list all of them, but examples are the U. S. Office of Education, National Science Foundation, U. S. Public Health Service, Office of Naval Research, College Entrance Examination Board, and many private foundations. An example of contract research in a school system is to be found in the Cooperative Research Project on the mentally retarded but educable child; the study is being conducted over a three-year period by the Chicago Board of Education. Out of this project are developing a number of evaluation instruments and rating devices which have been devised for use with educable mentally handicapped children but which will also have applicability for average and above-average children as well.

Interest in studying the predictive value of high-school indexes for college achievement continued from the previous period reviewed. Edward O. Swanson of the Student Counseling Bureau of the University of Minnesota reported in a personal communication to the author that he had studied simple and multiple correlations of such indexes with freshman grades for nearly all the Minnesota colleges, recommending the use of either regression equations or expectancy tables for practical interpretation of the results. Bennett, Seashore, and Wesman (8) reported a seven-year follow-up of high-school students tested on the *Differential Aptitude Tests* which showed certain profile differences among persons entering diverse careers.

Cureton and others (13) produced a specimen set of an abbreviated demonstration aptitude test battery, designed solely for instructional purposes. Included in the study kit are test booklets for two parallel forms, four different types of scoring keys, and a manual. The kit is inexpensive, is not confidential, and can fulfill a need for demonstration material in measurement classes where one is faced with problems of individual materials, security of tests, and testing time limits.

Larson and McCreary (46) surveyed testing practices in California public secondary schools. They found that the schools had a general

testing pattern and that although the test results were generally available to colleges, there were few requests for such information on the part of colleges.

Manuel (52) discussed selective admissions in public colleges and pointed to difficulties caused by the wide range of applicants' abilities and the lack of college resources. He cited the experience of The University of Texas in this respect. At the University of Kansas, Smith (72) studied the achievement of students in relation to certain cutting scores on admission tests, and Yocum and Anderson (88) studied the achievement of a group of mentally superior students.

Stuit, Helmstadter, and Frederiksen (75) surveyed the problems of college evaluation and pointed up the need for (a) evaluation instruments which more clearly emphasize important educational objectives; (b) better normative data for new instruments; and (c) better understanding of how to organize, instruct, and handle a small group such as an evaluation committee.

The mushrooming of electronic computer installations—some 1200 of them at last count by Wrigley (87)—portends powerful impetus to test development by making feasible kinds of research studies otherwise impractical, such as those in configural scoring of test items, pattern analysis of total scores, large-scale correlational analyses, and factor analyses.

Developments in Testing Programs

Several worthwhile developments in measurement and evaluation programs of school systems were noted by way of personal communications (which except for the first are not cited in the bibliography). Jackson of the Dearborn (Michigan) public schools reported (a) the use of item analysis by teachers and committees to improve their own test construction, including departmental achievement tests; and (b) the use of test-retest procedures to determine gains over a school year (42). James C. Adell of the Cleveland (Ohio) schools reported that practice tests to insure that pupils are familiar with standardized answer sheets are begun at the kindergarten level. Carleton B. Shay of the Santa Monica (California) high school emphasized the need for local norms and described the use of approximate norms which were especially helpful with atypical groups. Warren Findley, of the Atlanta (Georgia) schools wrote that the ratio of Mental Age to Chronological Age obtained from a group test is called the PLR (Probable Learning Rate) to distinguish it from the IQ as obtained from an individual test in a clinical setting.

The initial issue of the *NCMUE Newsletter* of the National Council on Measurements Used in Education (56) summarized a number of developments in testing programs. The New Bedford (Massachusetts) public schools reported the use of student help to solve the clerical problem of recording and interpreting test results. The Portland (Maine) public schools used a public address system to administer achievement and in-

telligence tests in three junior high schools. In Pinellas County (Florida) the testing leaders of the 10 largest counties in the state were invited to a one-day meeting to explore mutual problems.

The problems of planning, setting up, maintaining, and evaluating testing programs were considered by a number of workers among whom were Dobbin (19), Durost (25), Educational Testing Service (27), Elliott (28), Engelhart (29), Landy (45), Lennon (48, 49), Rankin (64), Seashore (69), Weitz (80), and Womer (84).

North (62) surveyed the testing programs of the public-school members of the Educational Records Bureau. He found that all members had a systematic testing program in operation and that tests of a wide variety were being used. Nearly all the members based their testing programs on stated educational objectives, recorded test results on a uniform cumulative record, and made results available to counselors and teachers. In a majority of the cases, members indicated that special appraisal devices were used; achievement test results were interpreted to the board of education; results were made available to parents and students under certain circumstances; inservice training programs were held to familiarize teachers with the interpretation and use of test results; and testing programs had resulted in improvements in guidance, pupil programing, instruction, curriculum development, and grouping procedures. In about half the cases it was indicated that teachers helped to score the tests; in very few cases were the tests scored exclusively by teachers. In only a few cases was it reported that the testing program was based in part upon locally constructed tests.

Harvey (41) inquired into the uses and practices in 296 institutions participating in the Graduate Record Examination Institutional Testing Program for 1955-56; he found (a) an increase in the number of institutions, and (b) the tendency to use the tests at a wider range of educational levels and for a greater variety of purposes, which were described in detail.

Testing played an important part in two new programs inaugurated by Educational Testing Service during the period (26). They were the Teacher Education Examination Program and the Sponsored Scholarship Services. One of the largest users of the latter program, the National Merit Scholarship Corporation, described the Merit Scholarship Program in its own brochure (57).

The status of the General Educational Development Testing Program and the results of its use were reported in a comprehensive brochure (3). The research activities of the College Entrance Examination Board were reported for a five-year period by Fishman (31).

Davis (14) reported upon a centralized testing and guidance service of the Board of Regents of the University System of Georgia, which is comprised of 15 tax-supported institutions of higher learning. The program has as its primary purpose the improvement of system-wide selection and counseling procedures. Technical matters relating to testing and re-

search methodology were placed in the hands of specialists, and the service acts in an advisory capacity, member institutions retaining their autonomy in formulating their own admissions policies. Extensive data on norms and validities are provided college counselors and high-school principals throughout the state.

There was a trend toward a greater interest of various religious groups in testing. A number of dissertations were written on selection tests for seminarians. Kling (44) reported a research study of the predictive value of psychological tests in the case of the Christian ministry. Resulting from the initial phase of the study were a criterion instrument which shows promise for future phases of the study and a bibliography on testing as related to the ministry (55). As an example of a broadening trend, one denomination held its third annual "career clinic" for high-school students throughout the state on the campus of a co-operating university. Another denomination published a tests and measurements manual for its school system (50).

Sources of Information on Testing

Excellent resources are to be found in several issues of the REVIEW which have contributed to the improvement of tests, to test usage, and to improved statistical methodology. The February 1956 issue entitled "Educational and Psychological Testing" covered the period from 1953 to 1956 (4). The June 1956 issue (6) was a special one entitled "Twenty-Five Years of Educational Research"; special note should be taken of the chapter in that issue entitled "Educational Measurements" by Wrightstone and others (85). Another chapter in the same issue with implications for measurement and evaluation was the one entitled "Methods of Research" by Walker (78). The December 1957 issue (5) entitled "Methodology of Educational Research" covered the period from 1953 to 1956 and included a chapter entitled "Research Tools: Scaling and Measurement Theory" by Messick and Abelson (54).

Several new textbooks on testing and evaluation appeared during the period. Outstanding among these was *Introduction to Educational Measurement* by Noll (59). Also noteworthy but of more specialized interest were *Specimen Objective Test Items* by Gerberich (38) and *Evaluation in the Basic College at Michigan State University* by Dressel (22).

Other new texts include *Measurement and Evaluation in Education* by Bradfield and Moredock (10), *Evaluation Techniques for Classroom Teachers* by Baron and Bernard (7), *Constructing Evaluation Instruments* by Furst (37), *Evaluating Student Progress in the Secondary School* by Schwartz and others (68), *Measurement and Evaluation for Secondary School Teachers* by Torgerson and Adams (76), and *Evaluation in Modern Education* by Wrightstone, Justman, and Robbins (86).

Recent revisions of texts include *Theory and Practice of Psychological Testing* by Freeman (36) and *Evaluation and the Elementary Curriculum*

by Shane and McSwain (71). The *Taxonomy of Educational Objectives* (9), previously published in a preliminary edition, appeared in its final form.

Several short, paperback books appeared. Notable among these was *Essentials of Educational Evaluation* by Wandt and Brown (79), designed for a unit on testing rather than a full semester course. Those who need guidance material to orient students toward the proper "set" for testing will welcome *Taking a Test* by Manuel (53). A brief but well-done treatment of test principles appropriate for the inservice training of teachers, counselors, and administrators is to be found in *Measuring Pupil Achievement* by Lefever, Naslund, and Thorpe (47).

Bibliography

1. ALLEN, MARGARET E. "Status of Measurement Courses for Undergraduates in Teacher-Training Institutions." *13th Yearbook, National Council on Measurements Used in Education.* New York: the Council (Secy.: Robert D. North, 21 Audubon Avenue), 1956. p. 69-73.
2. ALLISON, ROGER B., JR., and HELMSTADTER, GERALD C. "Communicating Test Information in a Test Manual." *Journal of Counseling Psychology* 3: 64-66; Spring 1956.
3. AMERICAN COUNCIL ON EDUCATION, COMMMITEE ON EVALUATION OF THE TYLER FACT-FINDING STUDY. *A Study of the General Educational Development Testing Program.* Washington, D. C.: the Council, 1956. 72 p.
4. AMERICAN EDUCATIONAL RESEARCH ASSOCIATION. "Educational and Psychological Testing." *Review of Educational Research* 26: 5-109; February 1956.
5. AMERICAN EDUCATIONAL RESEARCH ASSOCIATION. "Methodology of Educational Research." *Review of Educational Research* 27: 427-547; December 1957.
6. AMERICAN EDUCATIONAL RESEARCH ASSOCIATION. "Twenty-Five Years of Educational Research." *Review of Educational Research* 26: 199-344; June 1956.
7. BARON, DENIS, and BERNARD, HAROLD W. *Evaluation Techniques for Classroom Teachers.* New York: McGraw-Hill Book Co., 1958. 297 p.
8. BENNETT, GEORGE K.; SEASHORE, HAROLD G.; and WESMAN, ALEXANDER, G. *The D. A. T.—A Seven-Year Follow-Up.* Test Service Bulletin No. 49. New York: Psychological Corporation, November 1955. 8 p.
9. BLOOM, BENJAMIN, editor. *Taxonomy of Educational Objectives.* New York: Longmans, Green and Co., 1956. 207 p.
10. BRADFIELD, JAMES M., and MOREDOCK, H. STEWART. *Measurement and Evaluation in Education.* New York: Macmillan Co., 1957. 509 p.
11. COLEMAN, WILLIAM. "Assisting Teachers in Using Test Results." *Personnel and Guidance Journal* 36: 38-40; September 1957.
12. CRONBACH, LEE J., and GLESER, GOLDINE C. *Psychological Tests and Personnel Decisions.* Urbana: University of Illinois Press, 1957. 165 p.
13. CURETON, EDWARD E., and OTHERS. *The Multi-Aptitude Test: Study Kit.* New York: Psychological Corporation, 1955.
14. DAVIS, JUNIUS A., editor. "Selective Admissions in the University System: The Problem and Search for Solution." *Proceedings of the Second Annual Admissions Conference Sponsored by the Regents of the University System of Georgia.* Atlanta: Regents of the University System of Georgia, 1957. 50 p. (Mimeo.)
15. DIEDERICH, PAUL B. "Exercise Writing in the Field of the Humanities." *Proceedings of the 1957 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1958. p. 36-44.
16. DIEDERICH, PAUL B. "Pitfalls in the Measurement of Gains in Achievement." *School Review* 64: 59-63; February 1956.
17. DIEDERICH, PAUL B. *Simple Test Analysis Procedures for Single Classes.* Princeton, N. J.: Educational Testing Service, 1957. 26 p.

18. DIEDERICH, PAUL B. *Simplified Measurement Techniques for Teachers.* Paper presented to the American Educational Research Association, February 1958. Princeton, N. J.: Educational Testing Service, 1958. 6 p.
19. DOBBIN, JOHN E. "Educational Testing Service: A Group Process Technique in Test Building." *14th Yearbook, National Council on Measurements Used in Education.* New York: the Council (Secy.: Robert D. North, 21 Audubon Avenue), 1957. p. 106-109.
20. DOPPELT, JEROME E. *How Accurate Is a Test Score?* Test Service Bulletin No. 50. New York: Psychological Corporation, June 1956. 4 p.
21. DOPPELT, JEROME E. *Watch Your Weights.* Test Service Bulletin No. 52. New York: Psychological Corporation, December 1957. 4 p.
22. DRESSEL, PAUL L. *Evaluation in the Basic College at Michigan State University.* New York: Harper and Brothers, 1958. 248 p.
23. DRESSEL, PAUL L., and HILL, WALKER H. *Critical Thinking: A Guide to Instruction and Evaluation.* East Lansing: Michigan State University, Board of Examiners, September 1955. 85 p. (Mimeo.)
24. DU MAS, FRANK M. *Manifest Structure Analysis.* Missoula: Montana State University Press, 1955. 193 p.
25. DUROST, WALTER N. *What Constitutes a Minimal Testing Program for Elementary and Junior High School.* Revised edition. Test Service Notebook No. 1. Yonkers-on-Hudson, N. Y.: World Book Co., 1956. 6 p.
26. EDUCATIONAL TESTING SERVICE. *Annual Report, 1956-1957.* Princeton, N. J.: the Service, 1957. 96 p.
27. EDUCATIONAL TESTING SERVICE. *Essential Characteristics of a Testing Program.* Evaluation and Advisory Service Series, No. 2. Princeton, N. J.: the Service, 1955. 11 p.
28. ELLIOTT, MERLE H. "Testing Programs in Elementary School Districts." *California Journal of Educational Research* 7: 195-200; November 1956.
29. ENGELHART, MAX D. "Evaluation Testing Program—Some Basic Considerations." *Chicago Schools Journal* 37: 74-79; November-December 1955.
30. ENGELHART, MAX D. "Exercise Writing in the Social Sciences." *Proceedings of the 1957 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1958. p. 57-67.
31. FISHMAN, JOSHUA A. *Research Activities of the College Entrance Examination Board—1952-1957.* New York: College Entrance Examination Board, 1957. 112 p. (Mimeo.)
32. FLANAGAN, JOHN C. "The Evaluation of Methods in Applied Psychology and the Problem of Criteria." *Occupational Psychology* 30: 1-8; January 1956.
33. FLANAGAN, JOHN C. *Manual for School Administrators and Supervisors: The Personal and Social Development Program.* Chicago: Science Research Associates, 1956. 32 p.
34. FLANAGAN, JOHN C. *Teacher's Guide for the Personal and Social Development Program.* Chicago: Science Research Associates, 1956. 63 p.
35. FLANAGAN, JOHN C.; PUMROY, SHIRLEY S.; and TUSKA, SHIRLEY A. *The Development of a Criterion for Validating Personality Tests.* Paper presented to the American Psychological Association, September 1956. Pittsburgh, Pa.: American Institute for Research (410 Amberson Avenue), 1956. 4 p. (Mimeo.)
36. FREEMAN, FRANK S. *Theory and Practice of Psychological Testing.* Revised edition. New York: Henry Holt and Co., 1955. 609 p.
37. FURST, EDWARD J. *Constructing Evaluation Instruments.* New York: Longmans, Green and Co., 1958. 334 p.
38. GERBERICH, J. RAYMOND. *Specimen Objective Test Items.* New York: Longmans, Green and Co., 1956. 436 p.
39. GUSTAD, JOHN W. "Helping Students Understand Test Information." *Proceedings of the 1955 Invitational Testing Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1956. p. 51-59.
40. HART, IRENE. *Using Stanines To Obtain Composite Scores Based on Test Data and Teachers' Ranks.* Test Service Bulletin No. 86. Yonkers-on-Hudson, N. Y.: World Book Co., 1957. 4 p.
41. HARVEY, PHILIP R. *The Use of the Graduate Record Examinations by Colleges and Universities.* Princeton, N. J.: Educational Testing Service, 1957. 36 p. (Mimeo.)

42. JACKSON, JOSEPH. *Annual Report of the Department of Testing and Instructional Research.* Bulletin No. 522. Dearborn, Mich.: Dearborn Public Schools, July 1957. 23 p.
43. KAVRUCK, SAMUEL. "Thirty-Three Years of Test Research: A Short History of Test Development in the U. S. Civil Service Commission." *American Psychologist* 11: 329-33; July 1956.
44. KLING, FREDERICK R. "A Study of Testing as Related to the Ministry." *Religious Education* 53: 243-48; May 1958.
45. LANDY, EDWARD. "The Guidance Director's Problems and Suggestions for the Test Specialist." *Proceedings of the 1957 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1958. p. 100-11.
46. LARSON, CARL A., and MCCREARY, WILLIAM H. "Testing Programs and Practices in California Public Secondary Schools." *California Journal of Secondary Education* 31: 389-401; November 1956.
47. LEFEVER, D. WELTY; NASLUND, ROBERT A; and THORPE, LOUIS P. *Measuring Pupil Achievement.* Practical Ideas in Education Booklet. Chicago: Science Research Associates, 1957. 47 p.
48. LENNON, ROGER T. "Discussion of the School Administrator's Problems." *Proceedings of the 1957 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1958. p. 94-99.
49. LENNON, ROGER T. *Testing in the Secondary School.* Test Service Notebook No. 20. Yonkers-on-Hudson, N. Y.: World Book Co., 1957. 4 p.
50. LUTHERAN EDUCATION ASSOCIATION. *Tests and Measurements in Lutheran Education.* Fourteenth Yearbook. River Forest, Ill.: the Association (7400 Augusta Street), 1957. 115 p.
51. MACALUSO, CHARLES J., and DAILEY, JOHN T. "Development and Application of Performance Standards for Naval Petty Officers." *American Psychologist* 13: 303-305; June 1958.
52. MANUEL, HERSCHEL T. "Aptitude Tests for College Admission." *14th Yearbook, National Council on Measurements Used in Education.* New York: the Council (Secy.: Robert D. North, 21 Audubon Avenue), 1957. p. 20-27.
53. MANUEL, HERSCHEL T. *Taking a Test.* Yonkers-on-Hudson, N. Y.: World Book Co., 1956. 77 p.
54. MESSICK, SAMUEL J., and ABELSON, ROBERT P. "Research Tools: Scaling and Measurement Theory." *Review of Educational Research* 27: 487-97; December 1957.
55. NATIONAL COUNCIL OF CHURCHES, DEPARTMENT OF THE MINISTRY. *A Bibliography on Testing as Related to the Ministry.* New York: the Council (297 Fourth Avenue), 1958. 14 p. (Mimeo.)
56. NATIONAL COUNCIL ON MEASUREMENTS USED IN EDUCATION. *NCMUE Newsletter* 1: 1-4; December 1957.
57. NATIONAL MERIT SCHOLARSHIP CORPORATION. *The Merit Scholarship Program.* Evanston, Ill.: the Corporation (1580 Sherman Avenue), 1958. 16 p.
58. NEDELSKY, LEO. "Exercise Writing in the Natural Sciences." *Proceedings of the 1957 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1958. p. 45-56.
59. NOLL, VICTOR H. *Introduction to Educational Measurement.* Boston: Houghton Mifflin Co., 1957. 437 p.
60. NOLL, VICTOR H. "Requirements in Educational Measurement for Prospective Teachers." *School and Society* 82: 88-90; September 17, 1955.
61. NORTH, ROBERT D. "The Experimental Use of a Biographical Inventory in Four Public High Schools." *1955 Fall Testing Program in Independent Schools and Supplementary Studies.* Bulletin No. 67. New York: Educational Records Bureau, 1956. p. 77-83.
62. NORTH, ROBERT D. "Testing Programs of Public School Members of the Educational Records Bureau—Report of a Questionnaire Survey." *1956 Achievement Testing Program in Independent Schools and Supplementary Studies.* Bulletin No. 68. New York: Educational Records Bureau, 1956. p. 76-90.
63. PERLOFF, ROBERT. *Increasing Test Utility Through Teacher-Identification of Criterion Components.* Paper presented to the regional convention of the American Educational Research Association, Cleveland, March 1958. Chicago: Science Research Associates, 1958. 6 p. (Mimeo.)

64. RANKIN, PAUL T. "The School Administrator's Problems for Testers." *Proceedings of the 1957 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1958. p. 86-93.
65. RIMOLDI, HORACIO J. A. *Problem Solving as a Process.* Chicago: Loyola University, June 1958. 17 p. (Mimeo.)
66. RIMOLDI, HORACIO J. A. "A Technique for the Study of Problem Solving." *Educational and Psychological Measurement* 15: 450-61; Winter 1955.
67. SCHUMACHER, CHARLES. *How To Make a "Content-Objectives" Test Analysis.* Bulletin on Classroom Testing No. 9. Minneapolis: University of Minnesota, Bureau of Institutional Research, 1957. 8 p.
68. SCHWARTZ, ALFRED L., and OTHERS. *Evaluating Student Progress in the Secondary School.* New York: Longmans, Green and Co., 1957. 434 p.
69. SEASHORE, HAROLD G. "Is It Time To Re-evaluate Your Testing Program?" *Schoolmen's Week Proceedings.* Philadelphia: University of Pennsylvania Press, 1956. p. 196-208.
70. SEASHORE, HAROLD G. "Tests as Aids to Administration and Counseling in Junior Colleges." *Junior College Journal* 26: 504-508; May 1956.
71. SHANE, HAROLD G., and MCSWAIN, E. T. *Evaluation and the Elementary Curriculum.* Revised edition. New York: Henry Holt and Co., 1958. 436 p.
72. SMITH, GEORGE B. "Who Would Be Eliminated? A Study of Selective Admission to College." *University of Kansas Publications* 7: 1-28; December 1956.
73. STECKLEIN, JOHN E. *How To Make an Item Analysis of an Objective Test.* Bulletin on Classroom Testing No. 8. Minneapolis: University of Minnesota, Bureau of Institutional Research, 1957. 8 p.
74. STEIN, MORRIS I. "Criteria of Nonintellectual Aspects of Personality." *Proceedings of the 1957 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1958. p. 21-29.
75. STUIT, DEWEY B.; HELMSTADTER, GERALD C.; and FREDERIKSEN, NORMAN. *Survey of College Evaluation Methods and Needs.* A report to the Carnegie Corporation, December 1956. Princeton, N. J.: Educational Testing Service, 1957. 225 p.
76. TORGERSON, THEODORE L., and ADAMS, GEORGIA S. *Measurement and Evaluation for the Secondary School Teacher.* New York: Dryden Press, 1956. 658 p.
77. TRAXLER, ARTHUR E. "Are the Professional Test-Makers Determining What We Teach?" *School Review* 66: 144-51; June 1958.
78. WALKER, HELEN M. "Methods of Research." *Review of Educational Research* 26: 323-43; June 1956.
79. WANDT, EDWIN, and BROWN, GERALD W. *Essentials of Educational Evaluation.* New York: Henry Holt and Co., 1957. 117 p.
80. WEITZ, HENRY. "Minimum Essentials for a Testing Program." *American School Board Journal* 135: 41-43; September 1957.
81. WESMAN, ALEXANDER G. *Aptitude, Intelligence, and Achievement.* Test Service Bulletin No. 51. New York: Psychological Corporation, December 1956. 4 p.
82. WESMAN, ALEXANDER G. "The Obligations of the Test User." *Proceedings of the 1955 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1956. p. 60-65.
83. WILSON, ROBERT C. "Improving Criteria for Complex Mental Processes." *Proceedings of the 1957 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1958. p. 13-20.
84. WOMER, FRANK B. "Initiating a Testing Program." *Elementary School Journal* 57: 193-97; January 1957.
85. WRIGHTSTONE, J. WAYNE, and OTHERS. "Educational Measurements." *Review of Educational Research* 26: 268-91; June 1956.
86. WRIGHTSTONE, J. WAYNE; JUSTMAN, JOSEPH; and ROBBINS, IRVING. *Evaluation in Modern Education.* New York: American Book Co., 1956. 481 p.
87. WRIGLEY, CHARLES. "Data Processing: Automation in Calculation." *Review of Educational Research* 27: 528-43; December 1957.
88. YOCUM, DALE M., and ANDERSON, KENNETH E. "A Study of Exceptional Students Who Entered the University of Kansas in the Fall of 1954." *University of Kansas Publications* 8: 1-40; November 1957.

CHAPTER II

Development and Applications of Tests of General Mental Ability

WILLARD G. WARRINGTON and JOE L. SAUPE

THIS review is concerned primarily with the literature since 1955. However, because the report by Stanley (71) in 1953 is the latest REVIEW chapter on this topic, a few references from the 1952-55 period are included. In eliminating potential references to meet space commitments, the review was made as comprehensive as possible and was restricted to studies most relevant for workers in education. Coverage in many areas may consequently appear thin, but the number of studies reviewed is not necessarily indicative of the total amount of work on a given test or topic. Many studies of more clinical than educational interest were purposely excluded. This chapter was further limited generally to single- or double-score tests of intelligence.

General Considerations

In his review Stanley (71) suggested that traditional group intelligence measures were being threatened by differential ability batteries, particularly the *Differential Aptitudes Test (DAT)*. Although some of the newer differential batteries may be experiencing deserved popularity, the ease with which references for this review were collected testifies that the development and use of the venerable tests of general mental ability or, more simply, intelligence, continue at the usual rate. While any conflict may have been more apparent than real, it nevertheless appears that publishers of differential batteries, having been unable to supplant single-score tests, have decided to recognize their value. For example, the Psychological Corporation (62) recently provided evidence of predictive validity and norms for the sum of *DAT* scores, VR and NA, used as a single measure of scholastic aptitude. Similarly, Science Research Associates (66) offered $2V + R$ as a basis for estimating IQ's from their *Primary Mental Abilities* tests and provided normative information for this procedure on a revised profile sheet.

Discussions of the nature of intelligence continue to appear and, as implied above, different conceptions of its nature possess concomitantly varying implications for its measurement. Somewhat polar positions were discussed by Burt (15) and by Guilford (33). Burt (15) traced the history of the concept of intelligence from Plato and Aristotle to the present, arguing that the hypothesis that intelligence is characterized by being (a) cognitive, (b) general, and (c) innate, has been developed

throughout history and is generally substantiated by modern statistical techniques. Guilford (33), on the other hand, in a progress report of his factorial study of intelligence, enumerated, classified, and discussed some 45 factors which have been identified with various degrees of certainty. He stated that *intelligence* is useful as a semipopular term, but that only single-factor scores can do complete justice in describing individuals. Significantly, both these positions are backed by factor-analytic evidence. It is the particular approach to factor analysis that differentiates them.

Another view of intelligence was offered by Fromm and Hartman (28) who attempted to separate intelligence from intelligence testing, using clinical cases to illustrate interdependence and interrelationships of emotions and intellect. Heim (37) also discussed intelligence tests from a clinical point of view supported by general observations in her work. Minor (57) presented a theory of intelligence as background to his discussion of the results of administering a 20-item vocabulary test to a well-selected national sample of 1500 individuals. His findings were discussed from the point of view of the maximum utilization of the intellectual resources of the U. S.; he deemed it possible to classify jobs and the labor force on a ladder of verbal ability.

The above references document the fact that even though tests of general mental ability are the most widely used tests, if not the most useful, controversy concerning the meaning of intelligence and, consequently, methods for measuring it remains. That interest in intelligence and intelligence testing continues is borne out by the allotting of almost an entire issue of the *British Journal of Educational Psychology* (13) to this topic. Articles in this issue were devoted to the definition of intelligence, the nature-nurture controversy, the study of errors made on intelligence test items, the stability of intelligence test performance, and the effects of practice and coaching.

Individual Tests: Verbal and Nonverbal

Because of the evident popularity of the *Wechsler* tests for research purposes, this section is organized as (a) *Wechsler* tests and (b) others. Although the revised *Stanford-Binet Intelligence Scales (S-B)* probably still receive considerable use, the number of researches concerned directly with these scales has fallen off. However, correspondence from the publisher indicates that a new revision made up of the best subtests of the 1937 scales is forthcoming. This revision can be expected to generate new research interest.

Guertin, Frank, and Rabin (32) presented the third of their reviews of research with the *Wechsler-Bellevue Intelligence Scale (W-B)*. Covering the 1950-55 period, they noted that there was an increase in the number of well-controlled studies, that interest in the use of the *W-B* as a test of general intelligence as opposed to its use as a diagnostic instrument

increased, and that interest in sex differences was evident for the first time. None of the 139 studies covered in that review is mentioned here. Examples of research with the *W-B* were the factor analysis reported by Davis (23) which suggested that the *W-B* subtests have complex factor patterns for the 11 factors isolated, and the study by Rubin-Rabson (63) which demonstrated that four *W-B* verbal subtests have discrepancies in order and degree of difficulty of items.

The *Wechsler Adult Intelligence Scale (WAIS)* (78), a revised and restandardized version of the original *W-B I* and *W-B II* scales, appeared in 1955 and has since stimulated much research effort in its behalf. Cohen (18) investigated the factorial structure of the *WAIS* at four different age levels from 18 through 60 and found that three major correlated factors appeared at each level and that a strong general factor accounted for over half of the total variance of all subtests.

Goolishian and Ramsay (30) compared the *WAIS* and *W-B I*, using two populations, and concluded that full scale IQ's were significantly lower on the *WAIS* than on *W-B I*. Cole and Weleba (19) gave both the *W-B I* and the *WAIS* to the same 46 college students and found a large practice effect on all three IQ scores. Dana (22) compared four verbal subtests for the above two scales and found that the degree of correlation, .69 to .93, was a function of the number of new items in the *WAIS*. Thus, it would seem that the equivalence of the *WAIS* and the *W-B* has not been unequivocally demonstrated for several types of populations.

There appears to be considerable interest in short forms of the *WAIS*. Using the standardization data, Doppelt (24) selected the pair of verbal subtests which correlated highest with total verbal score and the pair of performance subtests which correlated highest with total performance scores. Again using the standardization data, Maxwell (56) compared Whitmyre and Pishkin (80) applied both Doppelt's method and the prorating system to the test records of 100 psychiatric patients and found that both methods gave high correlations, .94 and .95, with full scale scores. Again using the standardization data, Maxwell (56) compared all possible combinations of subtests of the *WAIS* with 17 short forms of the *W-B* and found that the best *WAIS* short forms are different from the best *W-B* short forms.

McNemar (54) and Gwynne-Jones (34) discussed the importance of careful evaluation of difference scores on the *WAIS* to avoid overinterpretations of chance differences.

Research with the *Wechsler Intelligence Scale for Children (WISC)* continued. Using an interesting experimental design, Price and Thorne (61) investigated the equivalence of the *WISC* and *W-B I* for 40 11- and 40 14-year-old children. Their data indicated considerable lack of equivalence of the two tests, particularly in the performance scores. Holland (38), Harlow and others (35), and Arnold and Wagner (5) compared performance on the *WISC* and the *S-B* and found that *S-B* scores

correlated higher with full scale and verbal scores than with performance scores. They concluded that the *WISC* is a reasonably valid measure of intelligence in the age range six to 14 years.

Other individual intelligence tests were also the subject of numerous research reports. In particular, interest in Raven's *Progressive Matrices (PM)* increased. Green and Ewert (31) presented normative data for the *PM* under group administration by slides to 1214 Rochester, Minnesota, children, aged six to 12. Sperrazzo and Wilkins (69) repeated this norming study for a St. Louis, Missouri, population and found lower correlations, .23 to .40, between the *PM* and other measures of intelligence than did Green and Ewert (31). Bolton (12) reported considerable success in using the *PM* for testing non-English-speaking children. With respect to the *Colored Progressive Matrices (CPM)*, Martin and Wiechers (55) found a high correlation, .91, between the *CPM* and *WISC*, whereas Stacey and Carleton (70) reported a considerably lower correlation of .55 for a similar population.

The *Columbia Mental Maturity Scale (CMMS)* (14) which appeared in 1953 attracted some research attention. Testing 70 fourth-grade children, Barratt (6) found that the *CMMS* correlated .61 with the *WISC* and .58 with the *PM*. French and Worcester (27) compared *S-B* and *CMMS* scores for 41 normal and 90 retarded six- to 12-year-olds and found a correlation of .67 between the two tests for both groups; they also found that the *CMMS* overestimated the mental ability of the poorer pupils.

In other developments in this general area, Porteus (60) summarized research and developments concerning his *Maze Test*. Copple (21) proposed a novel oral sentence-completion technique as a measure of intelligence. The report by Armitage and others (4) demonstrated that attempts to use the *Rorschach* as an effective intelligence test continued to show inconclusive results.

Performance Tests

Only limited work appears to have been done in this area. Orgel and Dreger (59) compared the *Arthur Adaptation of the Leiter Performance Scale (AALPS)* and the *S-B*, Form L, and concluded that the *AALPS* was valuable for appraising the child with a verbal handicap. Levinson (45) reported a reliability of .88 for the *Knox Cube Backward Test (KCB)* and a correlation of .60 for the *S-B* and the *KCB*. Jones and Rich (40) found that the *Goodenough Draw-a-Man Test* gave a quick and reasonably valid estimate of intelligence in an aged adult population.

Group Tests

A considerable amount of valuable unpublished information concerning presently available group tests of mental ability was obtained from the major test publishers. While only a small amount of this information can

be presented here, it is nevertheless evident that testing in this area is still in an active developmental stage.

The demise of the *ACE Psychological Examination* was probably partly responsible for the appearance of several new, well-developed tests (11, 25, 65) designed to predict success at the college level. The *Lorge-Thorndike Intelligence Tests* (47) and a revised edition of the *Henmon-Nelson Tests of Mental Ability* (43) also appeared. Both these tests seem to have exceptionally good normative data and cover all grades from early elementary through high school.

Group tests continued to be the subject of much reported research. Sheldon and Manolakes (67) compared the *California Test of Mental Maturity, S-Form (CTMM)* and the *S-B*, Form L, for 422 first- to sixth-graders and found no significant differences between mean IQ's. However, nearly half the pairs of scores differed by more than 10 IQ points. Altus (1) compared the verbal and nonverbal parts of the *CTMM* and the *WISC*, finding *r*'s of .71 for verbal and .67 for nonverbal scores. A 107-item bibliography (16) prepared by the California Test Bureau summarizes investigations involving the *CTMM* from 1935 to 1955.

Justman and Wrightstone (41) examined the scores of 1698 eighth-grade pupils on the *Pintner Intermediate Test* and the *Henmon-Nelson* and reported that for group appraisal the two tests were interchangeable even though the *H-N* IQ's tended to be somewhat above *Pintner* IQ's at the low-ability level and below them at the high level.

"Cultural bias" in intelligence tests continued to attract attention. The *Purdue Non-Language Test* (75) represents an example of a new test that was advertised as "culture fair." The items of this test require subjects to select unique elements from sets of geometric designs.

The *Davis-Eells Games (D-E)* were studied in several contexts. Love and Beach (48) administered the *D-E* to 579 third- and fourth-grade pupils who had also taken the *Kuhlmann-Anderson (K-A)* or the *CTMM*. They reported correlations of .53 and .60 between *D-E* and *K-A* and *CTMM* scores, respectively. Tate and Voss (74) investigated differences in race, residence, and sex of some 1200 fourth-, fifth-, and sixth-graders on *D-E* and *CTMM*. The two tests discriminated equally between races, but *D-E* discriminated more sharply between rural and urban pupils; only *D-E* items by Tate and Voss (74) provided meager support for the claims of the test's designers. The studies reported by Altus (2) and by Coleman and Ward (20) provided further generally negative evidence concerning the "fairness" of the *D-E* to lower-class children. The conclusions of Ludlow (49) in his summary of three *D-E* studies are still appropriate. Ludlow indicated that although the novel approach to test construction used in the *D-E* was to be commended, the research evidence so far reported did not provide conclusive evidence on the test's validity and that more study of the test was needed before it could be recommended as an operational instrument.

Applications of Tests of Mental Ability

The applications of intelligence tests reviewed here represent a very limited sample. More complete treatments in some major areas of application can be found in chapters by Pinneau and Jones and by Jones in the December 1958 REVIEW.

Particularly important during the past few years and very likely to be more important during the immediate future is the problem of predicting academic success. Stroud, Blommers, and Lauber (73) made a correlational analysis of the comparative value of the *WISC*, *S-B*, and *Iowa Tests of Basic Skills* as predictors of academic success in grades 3 to 6. Multiple R's of .60 to .75 were reported for various combinations of subtests. They made the provocative point that the population of pupils (referrals to school psychologists) on which their study was based, was *the* relevant population for the study of individual tests in school settings. For their sample the evidence indicated considerable predictive power for the intelligence tests. Wellman (79) reported that the *Otis Quick Scoring Mental Maturity Test* predicted ninth- and tenth-grade achievement better than any of the *SRA Primary Mental Abilities (PMA)* subtests or *PMA* total but that selected subtests of the *PMA* added significantly to the multiple predictive power of the combination. Russell (64) found that the *S-B*, Form L, was a better predictor of reading progress during first grade than the *Davis-Eells*. Barratt and Baumgarten (7) reported no differences between the *S-B* and the *WISC* in predicting reading or arithmetic achievement for achievers and nonachievers in grades 4 to 6. Jackson (39) investigated the effectiveness of several tests, including the *ACE*, for predicting success of college freshmen. He found that these tests had considerable predictive power but predicted better for women than for men. Klugman (42) used two tests, the *CTMM* and the *ACE*, to predict the success of 151 student nurses. He reported no difference in verbal but significant differences in the nonverbal area, *ACE* being a better predictor. In a summary report Lennon and Schutz (44) listed 479 correlations between several common group intelligence tests and various group achievement tests from unpublished studies during the period 1940 to 1956. Correlations reported range from .26 to .86 with a median *r* of .65. As in the past, then, and as might be predicted for the future, tests of general mental ability appeared to be most useful for predicting academic success at all levels. Such predictions, however, even when assisted by tests of other types, are not perfect.

Work in infant testing seemed somewhat limited for the review period. Cavanaugh and others (17) reported data showing that the *Cattell Infant Intelligence Scale* was a poor predictor of intelligence when administered at the age of six months. However, Simon and Bass (68) presented the clinician's point of view in arguing that rejection of the validity of infant testing seemed premature. They presented data from situations in

which test results were improved considerably by the use of clinical judgment.

Developmental and longitudinal studies of intelligence continued to receive considerable research attention. Bayley (8, 9, 10), in particular, reported extensively the results of repeated testing of the same sample of persons and presented an age curve of intelligence from birth to 50 years or older. She emphasized the complex nature of intellectual abilities and the difficulties inherent in the interpretation of measures of adult intelligence. Watts (77) reported that grammar-school girls, when tested with the same test annually for eight years, improved up to at least seven testings. She concluded that most of this gain was probably due to practice effect rather than increase in age. Tozer and Larwood (76) tested students at the beginning and at the end of their university degree courses and found statistical gains that were unrelated to age, sex, or course of study taken. Gehman and Matyas (29) administered the *WISC* and the *S-B* to 60 fifth-graders and again when these pupils were in the ninth grade, and reported that both tests yielded relatively constant IQ's over this period.

Consistent with the concern shown for the problems of aging, researches investigating the relationship of aging and intelligence were in evidence. Strother, Schaie, and Horst (72) reported test data for a sample of college graduates in the 70-84 age bracket. Their data showed that nonverbal abilities declined with advancing age much more rapidly than did verbal abilities. Lorge (46) summarized research relating aging and intelligence and concluded that intelligence and learning are maintained without significant decrease throughout early and middle maturity. This conclusion accords generally with the longitudinal studies mentioned above but is at variance with the more traditional idea of mental development that has been based largely on latitudinal studies, particularly those required in standardizing mental ability tests.

A variety of studies dealt with the relationship of intelligence to various physical and environmental factors. The comparison of Negro and white intelligence was investigated by Woods and Toal (81), using an analysis of variance design. They reported that Negroes scored higher on perceptual speed and accuracy and whites scored higher on culturally loaded items. McCord and Demerath (52) discussed previous studies of this same problem and presented data of their own that showed no significant relationship between race and intelligence when differences in socioeconomic status, parental education, and general home environment were statistically removed.

Anastasi (3) presented a comprehensive summary of research during the past 25 years on the relationship of intelligence and family size. She discussed the inconsistencies that have characterized research in this area and suggested explanations that might account for them. McArthur (50) attempted to show that intelligence tests are biased against upper classes

as well as lower classes. His results were in the predicted direction but were not conclusive. Estes (26) found that differences between upper and lower socioeconomic groups in *WISC* scores in the second grade were no longer apparent when the same pupils were tested in the fifth grade.

Two of a series of studies of the factorial composition of intelligence were reported by McCall (51) and McCormick (53). McCall (51) investigated sex differences in factor patterns, and McCormick (53) studied differences between high and low cognitive ability groups. Although generally similar factor patterns were found in both studies, differences were noted. For example, McCormick (53) reported that the "verbal" factor seemed relatively more important for the high ability group and the "cognitive" factor for the low, and that there was relatively more specificity of the primary factors for the high group.

Finally, two studies are cited as illustrative of the complexities that confound the measurement of mental ability. Mouly and Edgar (58) gave four well-known group tests to 164 ninth-graders and found considerable disparity in the IQ's for some students. They cautioned against too ready comparison and interpretation of IQ's obtained from different tests. Even more disturbing evidence, reported by Heim (36), was that test subjects tend to adapt their performance level to the level of difficulty of test items, i.e., do better on questions in a harder context than on the same questions in an easier context. Obviously there is considerable need for basic and integrative research in this area.

Bibliography

1. ALTUS, GRACE T. "Relationship Between Verbal and Non-Verbal Parts of the CTMM and WISC." *Journal of Consulting Psychology* 19: 143-44; April 1955.
2. ALTUS, GRACE T. "Some Correlates of the Davis-Eells Tests." *Journal of Consulting Psychology* 20: 227-32; June 1956.
3. ANASTASI, ANNE. "Intelligence and Family Size." *Psychological Bulletin* 53: 187-209; May 1956.
4. ARMITAGE, STEWART G., and OTHERS. "Predicting Intelligence from the Rorschach." *Journal of Consulting Psychology* 19: 321-29; October 1955.
5. ARNOLD, FRANK C., and WAGNER, WINIFRED K. "Comparison of Wechsler Children's Scale and Stanford-Binet Scores for Eight- and Nine-Year-Olds." *Journal of Experimental Education* 24: 91-94; September 1955.
6. BARRATT, ERNEST S. "The Relationship of the Progressive Matrices (1938) and the Columbia Mental Maturity Scale to the WISC." *Journal of Consulting Psychology* 20: 294-96; August 1956.
7. BARRATT, ERNEST S., and BAUMGARTEN, DORIS. "The Relationship of the WISC and Stanford-Binet to School Achievement." *Journal of Consulting Psychology* 21: 144; April 1957.
8. BAYLEY, NANCY. "Data on the Growth of Intelligence Between 16 and 21 Years as Measured by the Wechsler-Bellevue Scale." *Journal of Genetic Psychology* 90: 3-15; March 1957.
9. BAYLEY, NANCY. "A New Look at the Curve of Intelligence." *Proceedings of the 1956 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1957. p. 11-25.
10. BAYLEY, NANCY. "On the Growth of Intelligence." *American Psychologist* 10: 805-18; December 1955.
11. BENNETT, GEORGE K., and OTHERS. *College Qualification Tests (CQT).* New York: Psychological Corporation, 1957.

12. BOLTON, FLOYD B. "Experiments with the Raven's Progressive Matrices—1938." *Journal of Educational Research* 48: 629-33; April 1955.
13. BRITISH JOURNAL OF EDUCATIONAL PSYCHOLOGY. "Contributions to Intelligence Testing and the Theory of Intelligence." *British Journal of Educational Psychology* 27: 153-233; November 1957.
14. BURGOMEISTER, BESSIE; BLUM, LUCILLE H.; and LORGE, IRVING. *Columbia Mental Maturity Scale: Manual of Directions.* Yonkers-on-Hudson, N. Y.: World Book Co., 1953.
15. BURT, SIR CYRIL. "The Meaning and Assessment of Intelligence." *Eugenics Review* 47: 81-91; July 1955.
16. CALIFORNIA TEST BUREAU, DIVISION OF PROFESSIONAL SERVICES. *California Test of Mental Maturity: Summary of Investigations, Number Three.* Los Angeles: the Bureau, 1956. 30 p.
17. CAVANAUGH, MAXINE C., and OTHERS. "Prediction from the Cattell Infant Intelligence Scale." *Journal of Consulting Psychology* 21: 33-37; February 1957.
18. COHEN, JACOB. "The Factorial Structure of the WAIS Between Early Adulthood and Old Age." *Journal of Consulting Psychology* 21: 283-90; August 1957.
19. COLE, DAVID, and WELEBA, LOIS E. "Comparison Data on the Wechsler-Bellevue and the WAIS." *Journal of Clinical Psychology* 12: 198-99; April 1956.
20. COLEMAN, WILLIAM, and WARD, ANNIE W. "A Comparison of Davis-Eells and Kuhlmann-Finch Scores of Children from High and Low Socio-Economic Status." *Journal of Educational Psychology* 46: 465-69; December 1955.
21. COPPLE, GEORGE E. "Effective Intelligence as Measured by an Unstructured Sentence-Completion Technique." *Journal of Consulting Psychology* 20: 357-60; October 1956.
22. DANA, RICHARD H. "A Comparison of Four Verbal Subtests on the Wechsler-Bellevue, Form I, and the WAIS." *Journal of Clinical Psychology* 13: 70-71; January 1957.
23. DAVIS, PAUL C. "A Factor Analysis of the Wechsler-Bellevue Scale." *Educational and Psychological Measurement* 16: 127-46; February 1956.
24. DOPPELT, JEROME E. "Estimating the Full Scale Score on the Wechsler Adult Intelligence Scale from Scores on Four Subtests." *Journal of Consulting Psychology* 20: 63-66; February 1956.
25. EDUCATIONAL TESTING SERVICE. *Cooperative School and College Ability Tests (SCAT). Examiner's Manual: First Supplement.* Princeton, N. J.: the Service, 1956. 11 p.
26. ESTES, BETSY W. "Influence of Socio-Economic Status on the Wechsler Intelligence Scale for Children: Addendum." *Journal of Consulting Psychology* 19: 225-26; June 1955.
27. FRENCH, JOSEPH, and WORCESTER, DEAN A. "A Critical Study of the Columbia Mental Maturity Scale." *Exceptional Children* 23: 111-13; December 1956.
28. FROMM, ERIKA, and HARTMAN, LENORE D. *Intelligence: A Dynamic Approach.* New York: Doubleday and Co., 1955. 52 p.
29. GEHMAN, ILA H., and MATYAS, RUDOLPH P. "Stability of the WISC and Binet Tests." *Journal of Consulting Psychology* 20: 150-52; April 1956.
30. GOOLISHIAN, HAROLD A., and RAMSAY, ROSE. "The Wechsler-Bellevue Form I and the WAIS: A Comparison." *Journal of Clinical Psychology* 12: 147-51; April 1956.
31. GREEN, MRS. MEREDITH W., and EWERT, JOSEPHINE C. "Normative Data on Progressive Matrices (1947)." *Journal of Consulting Psychology* 19: 139-42; April 1955.
32. GUERTIN, WILSON; FRANK, GEORGE H.; and RABIN, ALBERT I. "Research with the Wechsler-Bellevue Intelligence Scale: 1950-1955." *Psychological Bulletin* 53: 235-57; May 1956.
33. GUILFORD, JOY P. "The Structure of Intellect." *Psychological Bulletin* 53: 267-93; July 1956.
34. GWYNNE-JONES, H. "The Evaluation of the Significant Differences Between Scaled Scores on the WAIS: The Perpetuation of a Fallacy." *Journal of Consulting Psychology* 20: 319-20; August 1956.
35. HARLOW, JUSTIN E., and OTHERS. "Preliminary Study of the Comparison Between the Wechsler Intelligence Scale for Children and Form L of the Revised Stanford-Binet Scale at Three Age Levels." *Journal of Clinical Psychology* 13: 72-73; January 1957.

36. HEIM, ALICE W. "Adaptation to Level of Difficulty in Intelligence Testing." *British Journal of Psychology* 46: 211-24; August 1955.
37. HEIM, ALICE W. *The Appraisal of Intelligence.* London: Methuen and Co., 1954. 171 p.
38. HOLLAND, GLEN A. "A Comparison of the WISC and Stanford-Binet IQ's of Normal Children." *Journal of Consulting Psychology* 17: 147-52; April 1953.
39. JACKSON, ROBERT A. "Prediction of the Academic Success of College Freshmen." *Journal of Educational Psychology* 46: 296-301; May 1955.
40. JONES, ALLAN W., and RICH, THOMAS A. "The Goodenough Draw-a-Man Test as a Measure of Intelligence in Aged Adults." *Journal of Consulting Psychology* 21: 235-38; June 1957.
41. JUSTMAN, JOSEPH, and WRIGHTSTONE, J. WAYNE. "A Comparison of Pupil Functioning on the Pintner Intermediate Test and Henmon-Nelson Test of Mental Ability." *Educational and Psychological Measurement* 13: 102-109; Spring 1953.
42. KLUGMAN, SAMUEL F. "Agreement Between Two Tests as Predictors of College Success." *Personnel and Guidance Journal* 36: 255-58; December 1957.
43. LAMKE, TOM A., and NELSON, MARTIN J. *Henmon-Nelson Tests of Mental Ability.* Revised edition. Boston: Houghton Mifflin Co., 1957.
44. LENNON, ROGER T., and SCHUTZ, RICHARD E. *A Summary of Correlations Between Results of Certain Intelligence and Achievement Tests.* Test Service Note-Book, No. 18. New York: World Book Co., 1957. 4 p.
45. LEVINSON, BORIS M. "The Knox Cube Backward (KCB) as a Performance Test of General Ability." *Journal of Clinical Psychology* 12: 185-87; April 1956.
46. LORGE, IRVING. "Aging and Intelligence." *The Neurologic and Psychiatric Aspects of the Disorders of Aging.* Proceedings of the Association for Research in Nervous and Mental Diseases. Baltimore: Williams and Wilkins, 1956. p. 46-60.
47. LORGE, IRVING, and THORNDIKE, ROBERT L. *Lorge-Thorndike Intelligence Tests, Technical Manual.* Boston: Houghton Mifflin Co., 1957. 16 p.
48. LOVE, MARY I., and BEACH, SYLVIA. "Performance of Children on the Davis-Eells Games and Other Measures of Ability." *Journal of Consulting Psychology* 21: 29-32; February 1957.
49. LUDLOW, H. GLEN. "Some Recent Research on the Davis-Eells Games." *School and Society* 84: 146-48; October 1956.
50. MCARTHUR, CHARLES. "Upper-Class Intelligence as the Critical Case for a Theory of Middle Class Bias." *Journal of Counseling Psychology* 4: 23-30; Spring 1957.
51. MCCALL, JOHN R. *Sex Differences in Intelligence: A Comparative Factor Study.* Studies in Psychology and Psychiatry, Vol. 9, No. 3. Washington, D. C.: Catholic University of America Press, 1955. 65 p.
52. MCCORD, WILLIAM M., and DEMERATH, NICHOLAS J., III. "Negro Versus White Intelligence; A Continuing Controversy." *Harvard Educational Review* 28: 120-35; Spring 1958.
53. MCCORMICK, SISTER WILLIAM PAULINE. *Factors of Intelligence in High and Low Cognitive Ability Groups.* Washington, D. C.: Catholic University of America Press, 1954. 49 p.
54. MCNEMAR, QUINN. "On WAIS Difference Scores." *Journal of Consulting Psychology* 21: 239-40; June 1957.
55. MARTIN, ANTHONY W., and WIECHERS, JAMES E. "Raven's Colored Progressive Matrices and the Wechsler Intelligence Scale for Children." *Journal of Consulting Psychology* 18: 143-44; April 1954.
56. MAXWELL, EILEEN. "Validity of Abbreviated WAIS Scales." *Journal of Consulting Psychology* 21: 121-26; April 1957.
57. MINOR, JOHN B. *Intelligence in the United States.* New York: Springer Publishing Co., 1957. 180 p.
58. MOULY, GEORGE J., and EDGAR, SISTER MARY. "Equivalence of IQ's for Four Group Intelligence Tests." *Personnel and Guidance Journal* 36: 623-26; May 1958.
59. ORGEL, ARTHUR R., and DREGER, RALPH M. "A Comparative Study of the Arthur-Leiter and Stanford-Binet Intelligence Scales." *Journal of Genetic Psychology* 86: 359-65; June 1955.

60. PORTEUS, STANLEY D. "Porteus Maze Test Developments." *Perceptual and Motor Skills* 6: 135-42; Second Quarter 1956.
61. PRICE, JOHN R., and THORNE, GARETH D. "A Statistical Comparison of the WISC and Wechsler-Bellevue, Form I." *Journal of Consulting Psychology* 19: 479-82; December 1955.
62. PSYCHOLOGICAL CORPORATION. *VR + NA—An Index of Scholastic Ability; Norms and Validity: Supplement to the Manual of the Differential Aptitude Tests.* New York: Psychological Corporation, 1958. 4 p.
63. RUBIN-RABSON, GRACE. "Item Order and Difficulty in Four Verbal Subtests of the Bellevue-Wechsler Scale." *Journal of Genetic Psychology* 88: 167-74: June 1956.
64. RUSSELL, IVAN L. "The Davis-Eells Test and Reading Success in First Grade." *Journal of Educational Psychology* 47: 269-70; May 1956.
65. SCIENCE RESEARCH ASSOCIATES. *Tests of Educational Ability, Technical Supplement.* Chicago: Science Research Associates, 1957. 14 p.
66. SCIENCE RESEARCH ASSOCIATES. *Manual for the SRA Primary Mental Abilities.* Third edition. Chicago: Science Research Associates, 1958. 31 p.
67. SHELDON, WILLIAM D., and MANOLAKES, GEORGE. "A Comparison of the Stanford-Binet, Revised Form L, and the California Test of Mental Maturity (S-Form)." *Journal of Educational Psychology* 45: 499-504; December 1954.
68. SIMON, ABRAHAM J., and BASS, LIBBY G. "Toward a Validation of Infant Testing." *American Journal of Orthopsychiatry* 26: 340-50; April 1956.
69. SPERRAZZO, GERALD, and WILKINS, WALTER L. "Further Normative Data on the Progressive Matrices." *Journal of Consulting Psychology* 22: 35-37; February 1958.
70. STACEY, CHALMERS L., and CARLETON, FREDERICK O. "The Relationship Between Raven's Colored Progressive Matrices and Two Tests of General Intelligence." *Journal of Clinical Psychology* 11: 84-85; January 1955.
71. STANLEY, JULIAN C., JR. "Developments and Applications of Tests of General Mental Ability." *Review of Educational Research* 23: 11-32; February 1953.
72. STROTHER, CHARLES R.; SCHAIE, K. WARNER; and HORST, PAUL. "The Relationship Between Advanced Age and Mental Abilities." *Journal of Abnormal and Social Psychology* 55: 166-70; September 1957.
73. STROUD, JAMES B.; BLOMMERS, PAUL; and LAUBER, MARGARET. "Correlation Analysis of WISC and Achievement Tests." *Journal of Educational Psychology* 48: 18-26; January 1957.
74. TATE, MERLE W., and VOSS, CHARLOTTE E. "A Study of the Davis-Eells Tests of Intelligence." *Harvard Educational Review* 26: 374-87; Fall 1956.
75. TIFFIN, JOSEPH; GRUBNER, ALIN; and INABA, KAY. *Purdue Non-Language Test, Preliminary Manual.* Chicago: Science Research Associates, 1958. 4 p.
76. TOZER, A. H. D., and LARWOOD, H. J. C. "The Changes in Intelligence Test Score of Students Between the Beginning and End of Their University Courses." *British Journal of Educational Psychology* 28: 120-28; June 1958.
77. WATTS, KATHLEEN P. "Intelligence Test Performance from 11 to 18: A Study of Grammar School Girls." *British Journal of Educational Psychology* 28: 112-19; June 1958.
78. WECHSLER, DAVID. *Wechsler Adult Intelligence Scale (WAIS).* New York: Psychological Corporation, 1955.
79. WELLMAN, FRANK E. "Differential Prediction of High School Achievement Using Single Score and Multiple Factor Tests of Mental Maturity." *Personnel and Guidance Journal* 35: 512-17; April 1957.
80. WHITMYRE, JOHN W., and PISHKIN, VLADIMIR. "The Abbreviated Wechsler Adult Intelligence Scale in a Psychiatric Population." *Journal of Clinical Psychology* 14: 189-91; April 1958.
81. WOODS, WALTER A., and TOAL, ROBERT. "Subtest Disparity of Negro and White Groups Matched for IQ's on the Revised Beta Tests." *Journal of Consulting Psychology* 21: 136-38; April 1957.

CHAPTER III

Development and Applications of Tests of Intellectual and Special Aptitudes

J. P. GUILFORD, BENJAMIN FRUCHTER, and H. PAUL KELLEY

THE CONCEPTION of fundamental differential aptitudes and the basic research aimed at their discovery were not taken seriously until Thurstone proposed his theory of multiple factors and carried out the first comprehensive factor-analytic study from this point of view, a little more than two decades ago (82). Since that time there has been substantial development along this line in the further work of Thurstone and his associates and in the Air Force wartime research on classification tests (40). By the end of the war, about 25 primary mental abilities (not all of them in the general intellectual area) had been found by Thurstone's methods of analysis. Many of them were demonstrated to have some importance in the selection and classification of aircraft pilots, navigators, bombardiers, and other aircrew personnel.

Since the war, the highest concentration of research along the same lines has been in connection with the project on Aptitudes of High-Level Personnel, at the University of Southern California.[1] In this project, the prevailing technique has represented a wedding of factor analysis with experimental method. The kinds of tasks or tests have been varied systematically, both qualitatively and quantitatively, according to hypotheses generated concerning the existence of certain primary intellectual abilities and concerning their properties. The batteries of tests have been administered to military personnel who were entering upon courses of training that in most cases led to the status of commissioned officers. The intellectual abilities under investigation were included under the heuristic categories of reasoning, creative thinking, evaluation, planning, and problem solving. Since the initiation of the project in 1949, a dozen major factor-analytic studies have been carried out, some of which have been reported recently (5, 50).

One of the obvious consequences of these studies is the continued indication that human intellect is a very complex phenomenon. The possibility that there is a unitary trait of intelligence, at least at adult levels, grows more remote. This is not a necessary consequence of the use of multiple-factor analysis. The results of an analysis are determined by the intercorrelations of the test scores. When tests are varied sufficiently in kind, zero correlations are numerous. The strongest logical support for the belief in a general intellectual factor has been the assertion that

[1] Under contract N60nr-23810 with the Office of Naval Research.

tests of intellectual abilities universally intercorrelate positively. This assertion is definitely not consistent with the facts.

Another consequence has been the discovery and verification of primary abilities in new areas such as creative thinking and planning. This is partly attributable to the use of new varieties of tests, but especially to a willingness to utilize tests of the completion or open-end type, tests that even require some subjective judgment in scoring. It does not seem to be possible to measure some of the more obviously creative talents by means of multiple-choice tests or other types of tests in which responses are not produced by the examinee but are presented to him. One implication of this is that an overwhelming emphasis upon completely objective testing could have serious educational consequences. Achievement tests, particularly, embody educational objectives and implement an educational philosophy, expressed or unexpressed.

Perhaps the most significant consequences have been the implications of the intellectual factors (a) for the assessment of individuals, (b) for the education of children and youth, and (c) for an understanding of intellect itself. Since the first two of these follow from the third, the picture of the nature of intellect that grows out of the studies will be presented briefly. From this picture, other implications may be readily deduced.

The Structure of Intellect

With the growing numbers of primary abilities discovered, it became increasingly important to attempt to find some unifying principles that would make possible an easier comprehension of the total list. Attempts to classify the factors have proved to be moderately successful, and in the process some significant principles of organization have emerged. This would not have been possible without the available knowledge of a sufficient number and variety of factors. The result of these attempts has been called a "structure of intellect" (34, 36, 37, 38). Although the organization of intellectual factors in a unified system, like most general theories, will probably undergo many changes as new information accumulates, in its present form it has proven very helpful in guiding further factorial research, and it seems to offer concepts that will be useful to the experimental psychologist as well as to the educator. Following is the authors' summary of the present viewpoint (some authorities, e.g., Burt, would not agree with this analysis; see Chapter II of this REVIEW).

The first and most obvious principle regarding the structure of intellect is that primary abilities differ according to the kind of material or content dealt with by the individual. For a long time we have had the recognition of a distinction between verbal and nonverbal tests. There prove to be verbal and nonverbal factors of intellect. But the nonverbal category subdivides into *two* classes of abilities. There are abilities to deal with "figural" material (concrete, perceived forms, and properties) on the one hand, and abilities to deal with "symbolic" material (composed

of letters, numbers, and the like) on the other. In the verbal category are abilities for dealing with concepts or meanings; hence, the third class of factors has been called "semantic." There are parallel abilities for dealing with the three kinds of content—figural, symbolic, and semantic.

Within each of the three categories as to content, factors differ with respect to the kinds of operations performed on the material. There are basically five kinds of operations as indicated by five kinds of factors. One operation is that of cognition, which simply has to do with knowing information. We discover or recognize perceived objects and their properties, we discover or recognize symbolic units, such as words and other expressions, and we discover or recognize meanings. A second kind of operation is that of memory or retention. An individual's memory is not equally good for all kinds of material or all kinds of information.

The third and fourth kinds of operations have to do with productive thinking. Productive thinking is involved when from given information some other information is generated. But it makes a difference whether the conclusion or other outcome is a unique one that is essentially determined by the information given or whether the generated information can be varied or must be varied, alternative outcomes being not only possible but also sometimes demanded. The former pertains to convergent thinking, thinking that converges upon the unique consequence. The latter pertains to divergent thinking, thinking that goes searching, changes route, and yields multiple answers. It is in the divergent-thinking category that we find the abilities most clearly associated with creative performance—fluency of thinking, flexibility of thinking, and originality.

The fifth kind of operation is evaluation. We are perpetually checking and rechecking our information, our memories, and our productions, convergent or divergent. In this connection we make use of feedback information that helps us to arrive at decisions as to the correctness, goodness, appropriateness, or suitability of our cognitions, memories, and conclusions. There is a set of evaluative abilities parallel to the productive-thinking abilities, memory abilities, and cognitive abilities.

The third major principle of classification of the primary intellectual abilities is in terms of the kinds of products achieved by the different kinds of operations applied to the different kinds of contents. We are not certain, as yet, that the same list of kinds of products applies in the case of every kind of operation and every kind of content, but enough is known to suggest that this may be so.

Six kinds of products have been recognized, and each kind results from the various kinds of operations. The kinds of products are units of information, classes of units, relations between units, patterns or systems of information, transformations, and implications. A few examples will show how operations, contents, and products combine in connection with factors. We cognize units of information in figural form. We re-

member related (associated) units of information in semantic form (ideas). A flexible thinker readily transforms information that comes to him in symbolic form, which suggests that he might be indulging in mathematical thinking to produce or to arrive at new information.

A Comprehensive Theory of Intellect

With three kinds of content, five kinds of operations, and six kinds of products involved in intellectual performances, there should be 3 x 5 x 6, or 90, primary intellectual abilities. About 50 of the primary intellectual abilities are now known through factorial investigations. It might thus seem that more than half of the possible intellectual factors are known, but there are other considerations that suggest that more than 90 potential factors exist.

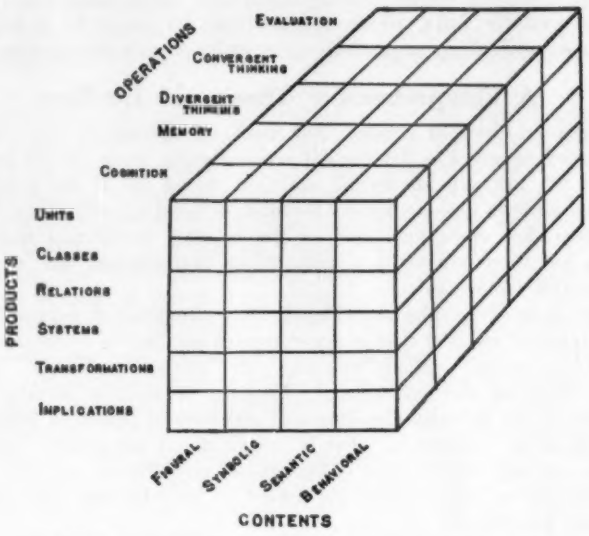
There is much empirical evidence of a nonfactorial nature concerning an area of intellect that is sometimes called "social intelligence" and more often recently called "empathy" or "empathic ability." As commonly conceived and investigated, this area of abilities pertains to the cognition of the thoughts, feelings, and attitudes of perceived individuals. If we think by analogy to what is known about recognized intellectual abilities, we may suppose that in the area of empathy we are dealing with a fourth kind of content, namely, a kind of material that may be designated as "behavioral."

Carrying the analogy further, we may hypothesize that the abilities for dealing with behavioral content are parallel to those already known (or predicted) in connection with the other kinds of content. The same operations theoretically apply so that besides abilities to cognize the behavior of others we have separate abilities for remembering behavior, for doing productive thinking about it, convergent and divergent, and for evaluating our cognitions and conclusions about it. Products of behavioral intellectual operations would be expected to fall in the same six categories—units, classes, relations, systems, transformations, and implications.

With the category of behavioral intelligence added, the comprehensive theory of human intellect, which has been elaborated in sketchy form, can be illustrated by means of a geometric model as in Figure I. The three principles of classification of the primary abilities are represented by the three dimensions of the cubical model. The order of the categories in each dimension is logical but not firmly fixed since we lack the criteria for establishing unique orders. There is little doubt that some of the intellectual factors have common linkages within individuals and would, therefore, exhibit some positive intercorrelations in populations. A reasonable, general prediction would be that correlations between factors are in direct proportion to proximity within the system when the orders of the categories are properly arranged.[2]

[2] Similar dimensional models for classifying primary traits have been found useful in the area of psychomotor abilities (39).

FIGURE I



THEORETICAL MODEL FOR THE COMPLETE "STRUCTURE OF INTELLECT"

Besides providing the basic variables along which individuals should be evaluated for various purposes, the structure of intellect suggests certain general implications for education. It may become popular, once again, to speak of education as development of the mind or of intellect. Knowing the intellectual abilities in all their variety and knowing their properties, we are in a much better position to suggest the course content and the procedures of instruction that should promote their improvement to the extent that they can be improved. In terms of learning theory, the implication definitely favors a cognitive bias in preference to the present stimulus-response bias. According to the cognitive view, the organism is an agent that discovers information, remembers information, and uses information in productive thinking and in evaluating any of its intellectual products. Such a view should have its consequences in the modification of philosophy of education with potentially far-reaching effects.

Tests of Differential Aptitudes (Aptitude Test Batteries)

Having looked at the theoretical aspects, we now turn to a review of the literature on applications of tests of special aptitudes for the period July 1955 to July 1958. *The Proceedings of the 1955 Invitational Con-*

ference on Testing Problems (2, 14, 29, 68) reported a series of papers on the use of multifactor ability test batteries in counseling and guidance. North (68) reported that counselors are making increased use of multifactor tests for differential predictions of academic and vocational success. These tests will become more useful in the school situation as reliability, validity, norms, and the theoretical framework of the factor scores are better determined. The general intelligence test still is useful in school counseling, however, especially at the elementary-school level. Anderson (2) described the *General Aptitude Test Battery (GATB)* and its use by the U. S. Employment Service. She pointed out that this battery provides the examinee with an Individual Aptitude Profile, covering nine aptitudes, as well as Occupational Aptitude Patterns, which indicate for which families of jobs the examinee is suited. Cureton (14) presented detailed descriptions of 16 batteries in tabular form. His presentation is of especial interest to counselors and others who must make some choice among the various published test batteries. French (29) concluded the session with a discussion of the logic of and assumptions underlying differential testing; every research worker and person interested in this problem should read this exposition carefully.

Michael (62) surveyed the field of differential testing with respect to the selection and placement of high-level personnel and considered a number of theoretical problems involved. Super (80) introduced a series of 10 articles on the use of multifactor test batteries in guidance, which appeared in the *Personnel and Guidance Journal*, by discussing the desiderata of guidance tests, the peculiarities of multifactor test batteries, the characteristics of available batteries, and implications for counseling. In each of the next eight articles one of the authors of a multifactor test battery presented the battery's origin, applicability, content, administration and scoring, norms, standardization and initial validation, reliability, validity, and use in counseling and selection. The test batteries reviewed were the *Differential Aptitude Tests (DAT)* (4), *General Aptitude Test Battery (GATB)* (18), *Guilford-Zimmerman Aptitude Survey* (35), *Holzinger-Crowder Uni-Factor Tests* (13), *Factored Aptitude Series of Business and Industrial Tests* (51), *Multiple Aptitude Tests* (76), *Flanagan Aptitude Classification Tests* (20), and *Tests of Primary Mental Abilities (PMA)* (83). Critical comments by Super were printed immediately following each article.

Super (79), in the final article of the series, provided a summary evaluation of each battery. He listed two batteries which he considered ready for use in counseling, four batteries he considered ready for research use only, and two he considered completely unsuitable for use at the present time.

Bennett (3) in a questionnaire study with a 59-percent return, reported that the *DAT* showed profile differences among high-school students who later entered various occupational and educational careers. Vineyard (85),

in another longitudinal study, examined the relationship between scores obtained early in high school and academic success in the freshman year of college. By means of multiple discriminant analysis Hall (42) determined similarities and differences between counselee groups; he was also able to determine a probability value for occupational and educational group membership for any subsequent examinee taking the *DAT* battery. Brayfield and Marsh (9) described the aptitude profiles of 50 farmers on the *DAT* battery. As compared with twelfth-grade students, these farmers excelled somewhat in *Mechanical Reasoning* but were average or below in the other aptitudes measured. Correlations with job performance measures were low, and relationship with job satisfaction was essentially zero.

Isaacson (47) used *GATB* scores and measures of general aptitude, interests, and personality traits to predict successful participation in a school's work experience program. Only one of the *GATB* scores—*General Intelligence*—was among the four measures found to be valid, the other three being measures of interest and personality. Retest performances on six tests of the *GATB* were investigated by Sorenson and Senior (77); one group was retested after four weeks, and a second group was retested after four years. Substantial gains were reported for most of the tests: For *Spatial Aptitude, Reasoning,* and *General Intelligence* scores the four-week gain was greater, while for *Verbal* and *Clerical Aptitude* scores the four-year gain was larger. A complication in interpreting the results was that the initial level of aptitude for the four-week group was lower than that for the four-year group.

A new manual for the *Guilford-Zimmerman Aptitude Survey* (41), containing additional information about research findings concerning these tests, was issued.

Mitchell (64) investigated the extent to which each of the four factor scores and a weighted composite score on the *Holzinger-Crowder Uni-Factor Tests* would predict high-school achievement in 14 communities. Nine of the 10 multiple correlations presented fell in the range .62 to .78. Bond and Clymer (7) reported that most of the *PMA* scores correlated significantly with reading ability, the *Space* test being the most notable exception.

The *Multi-Aptitude Test* (15) represents a multifactor test battery designed for purposes entirely different from those of the batteries considered previously. This battery of tests was developed specifically for demonstration, study, and practice use in tests and measurements courses, civic groups, layman conferences, and the like. French (28) analyzed a battery of pure-factor tests and criterion measures for a large sample of West Point plebes. The factor tests came out on the predicted factors with the exception of the *Space* tests which formed a cluster in the plane defined by the *Induction* and *Visualization* factors. The other factors were sufficiently independent to make it reasonable for them to be represented in a reference battery by a single pure test from each cluster.

Clerical and Mechanical Aptitudes

Several tests of clerical aptitude have been studied and validated. Hughes and McNamara (46) found the *Short Employment Tests* and the *General Clerical Tests* to be highly related, with a possible consequent saving in administration time in the selection of clerical applicants from use of the former. Prescott (72) validated the *Verbal Skills, Number Skills, Clerical Speed,* and *General Clerical Aptitude* scores of the *Turse Clerical Aptitudes Tests* against teachers' marks, teacher-made tests, and standardized tests on samples of commercial course students in two large high schools. The aptitude tests were administered at the beginning of the school year, and the criterion measures were obtained near the end of the school year. Interested in the problem of validity, Lawshe and Steinberg (55) made an important exploratory investigation of the demands of clerical jobs.

The results of two studies indicate the need for further development of the *Minnesota Clerical Test.* Longstaff and Beldo (57) demonstrated that practice effect increases an applicant's score when the currently used form is administered to him more than once. Kirkpatrick (52) found that an alert examinee who noticed that there were more discrepancies toward the end of the items in *Number Checking* and began checking from right to left had a considerable advantage. He recommended revision of the test to distribute the discrepancies more evenly.

McNamara and Hughes (58) administered *Letter-Digit Substitution* and *Name Checking* as well as a number of standardized tests for the selection of card punch operators. Moderate validities against course grades for trainees and somewhat higher validities for supervisors' ratings of employed operators were observed.

Crawford and Crawford (12) standardized an individual apparatus test of fine eye-hand co-ordination and developed percentile norms by sex for several adult job applicant groups.

Fleishman (23) gave a battery of 16 experimental manipulative tests to a sample of airmen engaged in mechanical work to determine whether manipulative tests would add to the validity of the operational battery for a final grade criterion in three technical training schools. He found that two paper-and-pencil tests, *Large Tapping,* and *Discrimination Reaction Time,* and an apparatus test, *Precision Steadiness,* added moderately to the validity of the battery then in use.

In a series of articles Fleishman (21, 24) and Fleishman and Hempel (25) reported the extensive exploratory work in psychomotor skills carried out in the USAF Aviation Psychology Program. Descriptions of the tests and the skills measured are given. Broad group factors of psychomotor skills were found which account for performance on a wide variety of psychomotor tasks. One of the factors was identified as *Integration* and was defined as the "ability to utilize a number of cues and activities quickly in making an integrated response." With extended

practice there was a considerable change in the pattern of abilities required to perform complex psychomotor tasks. Abilities which can be measured by printed tests, such as *visualization* and *spatial orientation*, accounted for most variance at early stages of practice, whereas aptitudes measured by apparatus tests such as *speed of arm movement* and *response orientation* were most prominent at later stages of proficiency.

Preference for use of principles over facts in solving problems on the *Balance Problems Test* was shown by Gaier (30) to add significantly to the multiple correlation of the best previously used predictors of final grades in the Air Force Airplane and Engines Mechanics Course. Gordon (32) demonstrated that scores on mechanical information tests should be corrected for amount of mechanical background as determined from biographical information of airmen since there is a negative relationship between background and success in training when the mechanical information score is held constant. Fifty-two items at a difficulty level appropriate for female naval recruits gave a more valid score than all 100 items of a mechanical test for male recruits in a study by Mollenkopf (65).

There have been several studies of the effect of training and background on the spatial visualization abilities. Blade and Watson (6) found that high scores on spatial visualization tests indicate an aptitude for engineering study but suggested that low scores may indicate only a lack of related past experience. Worsencroft (89) observed significant improvement in the spatial relations test scores of engineering students over the course of a year but scarcely any improvement for nonengineering students. He concluded that the improvement of the engineering students was due primarily to training in engineering drawing. Myers (67) observed a different result. Students at the U. S. Naval Academy with previous training in mechanical drawing had, on the average, test scores equal to the scores of those who had not had such training, but the group with the training received significantly better course grades in engineering drawing. Mendicino (61) found no significant differences between matched experimental and control tenth-grade student groups on *DAT Spatial Relations* and *Mechanical Reasoning* scores. The experimental group had taken vocational machine shop and mechanical drawing courses, whereas the control group had not. Michael and his associates (63) reviewed factor analytic studies of the spatial domain, gave definitions, and recommended reference tests for the three principal factors identified.

Artistic Aptitudes

Drake (17) published two forms of a musical aptitude test which yields two reliable scores: *Musical Memory* and *Rhythm*. The test is on microgroove records, and the two scores are reported to have low correlations with each other as well as with age and intelligence. The more difficult *Series B* of the *Seashore Measures of Musical Talent* was discontinued.

The 1956 *Manual* for *Series A* (75) contains references to reliability and validity studies and norms for grades 4 through 16.

Whittington (87) selected a group with musical background and ability to be compared with a nonmusical group. Some of the Wing (88) tests of musical intelligence discriminated between the two groups and were not correlated with tests of manual dexterity. Ottman (69) in a study of skills involved in sight singing in a college music group found that intelligence, language, reading, and *Seashore* measures were not significantly related, but that melodic modulation and hearing intervals with a harmonic background had the highest relationship to skill in sight singing.

Although there was no dearth of studies of the visual arts during the period covered by this review, none was considered worthy of inclusion.

Predicting Success in Professional Training (Professional School Aptitude Batteries)

Morici (66) studied 85 accounting graduates who had taken the *American Institute of Accountants Orientation Test.* For first-year accounting work the total score was most predictive, but for predicting subsequent achievement the *Q* score was the best.

In investigating the prediction of first-year grades at the Emory University Dental School by means of the *Aptitude Tests of the American Dental Association* and by the use of predental grades, Webb (86) reported the results of two studies. Zero-order validity coefficients were low, but the multiple correlations found were .43 and .50.

Jones and Case (49) reported information about an aptitude test battery developed by the Engineering Schools of the University of California at Los Angeles and Berkeley for use with lower-division applicants. While the separate validity coefficients were quite low, multiple correlations of about .50 were obtained. Kirkpatrick (53) investigated the efficiency of a battery of aptitude and personality measures for both the selection and placement of engineers. Harrison, Hunt, and Jackson (43) compared 240 mechanical engineers with the general population with respect to test norms on tests of mental ability. On every test the mean score for the engineers was well above that for the general population, but the engineers seemed to be no more superior on the engineering aptitude tests than on the tests of general intellectual ability. Moreover, the engineers did as well on the verbal tests as on such tests as *Space Relations, Mechanical Comprehension,* and *Abstract Thinking.* In an analysis of the specific activities of a large number of engineers, seven distinct areas were identified (19). Interest measures yielded the greatest differentiation among the specialties, next were tests of various abilities, and then came measures of certain personality traits. Boyce (8) tried various methods of combining scores to predict a dichotomous criterion of success in a cooperative engineering program. Gross quantitative methods did as well as refined techniques and showed less shrinkage on cross validation.

A wealth of information with regard to the *Law School Admission Test* is contained in the handbook prepared by Johnson, Olsen, and Winterbottom (48). In addition to discussing the test itself and techniques for using test scores in admissions, scholarship selection, counseling, comparing classes, and in relation to requests for deferment, a section is devoted to reporting the results from research studies. The six appendixes contain useful statistical and research information.

Gray, Duncan, and Davis (33) studied the validity of the *Iowa Legal Aptitude Test* as a predictor of first-year law grades. The obtained validity coefficients were significant but low. Martin (59) attempted to predict graduation from a law school and performance on state bar examinations from scores on a battery of entrance tests. The graduates were superior to the nongraduates on the entrance tests. Some parts of the *Iowa Legal Aptitude Test* contributed significantly to prediction.

Capps and DeCosta (11) sought to determine the extent to which *Graduate Record Examination* scores, *National Teacher Examinations* scores, and undergraduate grade-point averages were related to graduate-school success for students who were primarily teachers and teacher candidates. Multiple correlations of .57 to .59 were obtained by using various combinations of the predictor variables.

Das (16) presented a review of the literature on the selection of medical students. He concluded that so-called medical or professional aptitude tests may sometimes prove useful for predicting medical-school achievement but only to a limited extent. Stalnaker (78) presented mean scores on the *MCAT* for applicants accepted by medical schools and for those not accepted. Melton (60) studied 102 male premedical freshmen at the University of Minnesota. At the end of the freshman year the variables which discriminated between students admitted to medical school and other students were high-school rank, *ACE Psychological Examination*, and first-year honor-point ratio.

Hill (44) reported correlations of pharmacy grades with the *ACE Psychological Examination*, the *Ohio State Psychological Examination*, the *Purdue Mathematical Training Test*, and the *Iowa Chemistry Test*. A multiple correlation of .61 was obtained by using the first three tests; the inclusion of the chemistry test failed to add to this value.

Miscellaneous

In a review of the pertinent literature Patterson (70) pointed out that comparatively little has been done on the prediction of success in trade-school training although the problem is as important as the prediction of success in college work because of the number of people involved. The best predictors have been tests of verbal intelligence, mechanical information, and spatial ability. In another study Patterson (71) concluded that persistence in trade-school training could be predicted to a signifi-

cant degree with the battery he tried out, but that much improvement was still needed. French (27) administered an extensive battery of aptitude and interest measures in two vocational and technical high schools. There were a number of tests with validity coefficients suitable for comparative prediction of success in different shop courses; the prediction of occupational criteria was less satisfactory.

Lee (56) analyzed tests of general intellectual ability and of fundamental processes in mathematics at five grade levels. She found three factors indicating "an organization of mental abilities corresponding to the hypothetical structure of mathematical thinking upon which the ability tests were based." Hills (45) obtained the validities of eight cognitive factors for grades in calculus, college mathematics, and ratings by instructors. The verbal score of the *Junior Scholastic Aptitude Test* was shown by Traxler (84) to have a moderate validity for predicting achievement in first-year French and Latin.

Rimland (73) reported on the development and validation of new forms of the *NROTC Contract Student Selection Test*. Although the new forms were more reliable and of more appropriate difficulty for college freshmen, they were no more valid for first-semester grades. Allison (1) studied the relationships between test scores and training-course grades for recruits from different backgrounds. The regression of school grades on predictor variables was sufficiently different for the two groups for him to recommend separate selection procedures for them.

Lauer (54) found group paper-and-pencil tests to be somewhat more valid than certain psychophysical tests for selecting Army driver personnel. Fleishman (22) developed and adapted some auditory tests for the selection of radiotelegraphers; the new tests were more highly related to a rate-of-learning code score than were the aptitude tests then in use. A factor analysis by Fleishman, Roberts, and Friedman (26) indicated that the code learning criterion had loadings on *Speed of Closure*, *Auditory Rhythm Perception*, and *Auditory Perceptural Speed* factors.

Ghiselli (31) made an extensive survey of the validity of tests for training criteria and occupational proficiency. He concluded that performances in training and on the job involve very different patterns of abilities, and that training criteria are more predictable than job performance. Rusmore and Toorenaar (74) used cost-accounting procedures to demonstrate the considerable savings that could be realized from the selection of applicants for telephone operator training by means of a battery of three valid tests. Cantoni (10) found that grade-point average, *Kuhlmann-Anderson Intelligence Test* scores, and scores on the *Bell Adjustment Inventory* had moderate relationships to later occupational status for a group of male students.

As Thorndike (81) pointed out in the previous review of this topic, the "Validity Information Exchange" section of *Personnel Psychology* is a good source for validity data, especially for the *USES General Apti-*

tude Battery and other occupational tests. A validity report section for educational and psychological tests now appears in *Educational and Psychological Measurement.*

Bibliography

1. ALLISON, ROGER B., JR. "Differential Performance of Fleet and Recruit Personnel in Torpedoman's Mates School." *Journal of Applied Psychology* 39: 393-96; December 1955.
2. ANDERSON, PAULINE K. "The Use of the General Aptitude Test Battery in the Employment Service." *Proceedings of the 1955 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1956. p. 16-21.
3. BENNETT, GEORGE K. *The D.A.T.—A Seven Year Follow-Up.* Test Service Bulletin, No. 49. New York: Psychological Corporation, 1955. 8 p.
4. BENNETT, GEORGE K.; SEASHORE, HAROLD G.; and WESMAN, ALEXANDER G. "The Differential Aptitude Tests: An Overview." *Personnel and Guidance Journal* 35: 81-91; October 1956.
5. BERGER, RAYMOND M.; GUILFORD, JOY P.; and CHRISTENSEN, PAUL R. *A Factor-Analytic Study of Planning Abilities.* Psychological Monographs, No. 435. Washington, D. C.: American Psychological Association, 1957. 31 p.
6. BLADE, MARY F., and WATSON, WALTER S. *Increase in Spatial Visualization Test Scores During Engineering Study.* Psychological Monographs, No. 397. Washington, D. C.: American Psychological Association, 1955. 13 p.
7. BOND, GUY L., and CLYMER, THEODORE W. "Interrelationship of the SRA Primary Mental Abilities, Other Mental Characteristics, and Reading Ability." *Journal of Educational Research* 49: 131-36; October 1955.
8. BOYCE, JAMES E. *Comparison of Methods of Combining Scores To Predict Academic Success in a Co-operative Engineering Program.* Doctor's thesis. Lafayette, Ind.: Purdue University, 1955. 76 p. Abstract: *Dissertation Abstracts* 15: 2286; No. 11, 1955.
9. BRAYFIELD, ARTHUR H., and MARSH, MARY M. "Aptitudes, Interests and Personality Characteristics of Farmers." *Journal of Applied Psychology* 41: 98-103; April 1957.
10. CANTONI, LOUIS J. "High School Tests and Measurements as Predictors of Occupational Status." *Journal of Applied Psychology* 39: 253-55; August 1955.
11. CAPPS, MARIAN P., and DECOSTA, FRANK A. "Contributions of the Graduate Record Examinations and the National Teacher Examinations to the Prediction of Graduate School Success." *Journal of Educational Research* 50: 383-89; January 1957.
12. CRAWFORD, JOHN E., and CRAWFORD, DOROTHEA M. *Crawford Small Parts Dexterity Test.* New York: Psychological Corporation, 1956.
13. CROWDER, NORMAN A. "The Holzinger-Crowder Uni-Factor Tests." *Personnel and Guidance Journal* 35: 281-87; January 1957.
14. CURETON, EDWARD E. "Service Tests of Multiple Aptitudes." *Proceedings of the 1955 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1956. p. 22-39.
15. CURETON, EDWARD E., and OTHERS. *The Multi-Aptitude Test.* New York: Psychological Corporation, 1955.
16. DAS, R. C. "The Selection of Medical Students." *Occupational Psychology* 30: 27-42; January 1956.
17. DRAKE, RALEIGH M. *Drake Musical Aptitude Tests.* Chicago: Science Research Associates, 1954.
18. DVORAK, BEATRICE J. "The General Aptitude Test Battery." *Personnel and Guidance Journal* 35: 145-52; November 1956.
19. EDUCATIONAL TESTING SERVICE. *Annual Report to the Board of Trustees, 1954-1955.* Princeton, N. J.: the Service, 1956. 140 p.
20. FLANAGAN, JOHN C. "The Flanagan Aptitude Classification Tests." *Personnel and Guidance Journal* 35: 495-504; April 1957.
21. FLEISHMAN, EDWIN A. "A Comparative Study of Aptitude Patterns in Unskilled and Skilled Psychomotor Performances." *Journal of Applied Psychology* 41: 263-72; August 1957.

22. FLEISHMAN, EDWIN A. "Predicting Code Proficiency of Radiotelegraphers by Means of Aural Tests." *Journal of Applied Psychology* 39: 150-55; June 1955.
23. FLEISHMAN, EDWIN A. *Predicting Success in Certain Aircraft Maintenance Specialties by Means of Manipulative Tests.* U. S. Air Force Personnel and Training Research Center, Research Report 55-23. Lackland Air Force Base, Texas: Personnel and Training Research Center, September 1955. 30 p.
24. FLEISHMAN, EDWIN A. "Psychomotor Selection Tests: Research and Application in the United States Air Force." *Personnel Psychology* 9: 449-67; Winter 1956.
25. FLEISHMAN, EDWIN A., and HEMPEL, WALTER E., JR. "Factorial Analysis of Complex Psychomotor Performance and Related Skills." *Journal of Applied Psychology* 40: 96-104; April 1956.
26. FLEISHMAN, EDWIN A.; ROBERTS, MILLARD M.; and FRIEDMAN, MORTON P. "A Factor Analysis of Aptitude and Proficiency Measures in Radiotelegraphy." *Journal of Applied Psychology* 42: 129-35; April 1958.
27. FRENCH, JOHN W. *Educational and Occupational Validation of a High School Comparative Prediction Battery (Pilot Study).* Research Bulletin 56-1. Princeton, N. J.: Educational Testing Service, February 1956. 45 p.
28. FRENCH, JOHN W. "The Factorial Invariance of Pure-Factor Tests." *Journal of Educational Psychology* 48: 93-109; February 1957.
29. FRENCH, JOHN W. "The Logic of and Assumptions Underlying Differential Testing." *Proceedings of the 1955 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1956. p. 40-48.
30. GAIER, EUGENE L. "Technique of Problem Solving as a Predictor of Achievement in a Mechanics Course." *Journal of Applied Psychology* 39: 416-18; December 1955.
31. GHISELLI, EDWIN E. "The Measurement of Occupational Aptitude." *University of California Publications in Psychology* 8: 101-216; December 1955.
32. GORDON, MARY A. "Patterns of Mechanical Background and Aptitude." *Educational and Psychological Measurement* 17: 408-15; Autumn 1957.
33. GRAY, CARLINGFORD; DUNCAN, KASPAR T.; and DAVIS, JUNIUS A. "A Validation Study of the Iowa Legal Aptitude Test." *Educational and Psychological Measurement* 15: 499-501; Winter 1955.
34. GUILFORD, JOY P. "Les Dimensions de l'Intellect." *L'Analyse Factorielle et ses Applications.* (Edited by H. Laugier.) Paris: Centre National de la Recherche Scientifique, 1955. p. 55-77.
35. GUILFORD, JOY P. "The Guilford-Zimmerman Aptitude Survey." *Personnel and Guidance Journal* 35: 219-24; December 1956.
36. GUILFORD, JOY P. "New Frontiers of Testing in the Discovery and Development of Human Talent." *Seventh Annual Western Regional Conference on Testing Problems.* Los Angeles: Educational Testing Service, 1958. p. 20-32.
37. GUILFORD, JOY P. *A Revised Structure of Intellect.* Reports from the Psychological Laboratory, No. 19. Los Angeles: University of Southern California, April 1957. 27 p.
38. GUILFORD, JOY P. "The Structure of Intellect." *Psychological Bulletin* 53: 267-93; July 1956.
39. GUILFORD, JOY P. "A System of the Psychomotor Abilities." *American Journal of Psychology* 71: 164-74; March 1958.
40. GUILFORD, JOY P., and LACEY, JOHN I., editors. *Printed Classification Tests.* AAF Aviation Psychology Research Program Reports, No. 5. Washington, D. C.: Superintendent of Documents, Government Printing Office, 1947. 919 p.
41. GUILFORD, JOY P., and ZIMMERMAN, WAYNE S. *The Guilford-Zimmerman Aptitude Survey with Manual of Instructions and Interpretations.* Second edition. Beverly Hills, Calif.: Sheridan Supply Co., 1956.
42. HALL, ROBERT C. "Occupational Group Contrasts in Terms of the Differential Aptitude Tests: An Application of Multiple Discriminant Analysis." *Educational and Psychological Measurement* 17: 556-67; Winter 1957.
43. HARRISON, ROSS; HUNT, WINSLOW; and JACKSON, THEODORE A. "Profile of the Mechanical Engineer: I. Ability." *Personnel Psychology* 8: 219-34; Summer 1955.
44. HILL, SUZANNE D. "The Relationship Between Grades and a Predictive Test Battery in the School of Pharmacy of the George Washington University." *Journal of Applied Psychology* 41: 61-62; February 1957.

45. HILLS, JOHN R. "Factor-Analyzed Abilities and Success in College Mathematics." *Educational and Psychological Measurement* 17: 615-22; Winter 1957.
46. HUGHES, JOHN L., and MCNAMARA, WALTER J. "Relationship of Short Employment Tests and General Clerical Tests." *Personnel Psychology* 8: 331-37; Autumn 1955.
47. ISAACSON, LEE E. "Predicting Success in the Work Experience Program." *Personnel and Guidance Journal* 33: 270-73; January 1955.
48. JOHNSON, A. PEMBERTON; OLSEN, MARJORIE A.; and WINTERBOTTOM, JOHN A. *The Law School Admission Test and Suggestions for Its Use: A Handbook for Law School Deans and Admissions Officers.* Princeton, N. J.: Educational Testing Service, 1955. 148 p.
49. JONES, MARGARET H., and CASE, HARRY W. "The Validation of a New Aptitude Examination for Engineering Students." *Educational and Psychological Measurement* 15: 502-508; Winter 1955.
50. KETTNER, NORMAN W.; GUILFORD, JOY P.; and CHRISTENSEN, PAUL R. "A Factor-Analytic Investigation of the Factor Called General Reasoning." *Educational and Psychological Measurement* 16: 438-53; Winter 1956.
51. KING, JOSEPH E., JR. "Factored Aptitude Series of Business and Industrial Tests." *Personnel and Guidance Journal* 35: 351-58; February 1957.
52. KIRKPATRICK, DONALD L. "The Minnesota Clerical Test." *Personnel Psychology* 10: 53-54; Spring 1957.
53. KIRKPATRICK, JAMES J. "Validation of a Test Battery for the Selection and Placement of Engineers." *Personnel Psychology* 9: 211-27; Summer 1956.
54. LAUER, ALVAH R. "Comparison of Group Paper-and-Pencil Tests with Certain Psychophysical Tests for Measuring Driving Aptitude of Army Personnel." *Journal of Applied Psychology* 39: 318-21; October 1955.
55. LAWSHE, CHARLES H., and STEINBERG, MARTIN D. "Studies in Synthetic Validity: I. An Exploratory Investigation of Clerical Jobs." *Personnel Psychology* 8: 291-301; Autumn 1955.
56. LEE, DORIS M. "A Study of Specific Ability and Attainment in Mathematics." *British Journal of Educational Psychology* 25: 178-89; November 1955.
57. LONGSTAFF, HOWARD P., and BELDO, LESLIE A. "Practice Effect on the Minnesota Clerical Test When Alternate Forms Are Used." *Journal of Applied Psychology* 42: 109-11; April 1958.
58. MCNAMARA, WALTER J., and HUGHES, JOHN L. "The Selection of Card Punch Operators." *Personnel Psychology* 8: 417-27; Winter 1955.
59. MARTIN, RICHARD R. *An Investigation of the Effectiveness of an Entrance Test Battery for Predicting Success in Law School.* Doctor's thesis. Philadelphia: Temple University, 1954. 207 p. Abstract: *Dissertation Abstracts* 16: 575-76; No. 3, 1956.
60. MELTON, RICHARD S. "Differentiation of Successful and Unsuccessful Premedical Students." *Journal of Applied Psychology* 39: 397-400; December 1955.
61. MENDICINO, LORENZO. "Mechanical Reasoning and Space Perception: Native Capacity or Experience." *Personnel and Guidance Journal* 36: 335-38; January 1958.
62. MICHAEL, WILLIAM B. "Differential Testing of High-Level Personnel." *Educational and Psychological Measurement* 17: 475-90; Winter 1957.
63. MICHAEL, WILLIAM B., and OTHERS. "The Description of Spatial Visualization Abilities." *Educational and Psychological Measurement* 17: 185-99; Summer 1957.
64. MITCHELL, BLYTHE C. "The Relation of High School Achievement to the Abilities Measured by the Holzinger-Crowder Uni-Factor Tests." *Educational and Psychological Measurement* 15: 487-90; Winter 1955.
65. MOLLENKOPF, WILLIAM G. "An Easier 'Male' Mechanical Test for Use with Women." *Journal of Applied Psychology* 41: 340-43; October 1957.
66. MORICI, ANTHONY R. "Relation Between the Scores on the A.I.A. Orientation Test with the A.I.A. Elementary, Advanced Accounting Tests and Accounting Grades." *Journal of Educational Research* 51: 549-52; March 1958.
67. MYERS, CHARLES T. *The Effects of Training in Mechanical Drawing on Spatial Relations Test Scores as Predictors of Engineering Drawing Grades.* Research Bulletin 58-4. Princeton, N. J.: Educational Testing Service, March 1958. 10 p.

68. NORTH, ROBERT D. "The Use of Multi-Factor Aptitude Tests in School Counseling." *Proceedings of the 1955 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1956. p. 11-15.
69. OTTMAN, ROBERT W. *A Statistical Investigation of the Influence of Selected Factors on the Skill of Sight-Singing.* Doctor's thesis. Denton: North Texas State College, 1956. 286 p. Abstract: *Dissertation Abstracts* 16: 763; No. 4, 1956.
70. PATTERSON, CECIL H. "Predicting Success in Trade and Vocational School Courses: Review of the Literature." *Educational and Psychological Measurement* 16: 352-400; Autumn 1956.
71. PATTERSON, CECIL H. "The Prediction of Attrition in Trade School Courses." *Journal of Applied Psychology* 40: 154-58; June 1956.
72. PRESCOTT, GEORGE A. "Prediction of Achievement in Commercial Subjects." *Educational and Psychological Measurement* 15: 491-92; Winter 1955.
73. RIMLAND, BERNARD. *The Development and Validation of Forms 1 and 2 of the NROTC (Contract) Student Selection Test.* U. S. N. Bureau of Naval Personnel Technical Bulletin, No. 55-17. San Diego, Calif.: U. S. Naval Personnel Research Field Activity, August 1955. 41 p.
74. RUSMORE, JAY T., and TOORENAAR, GERARD J. "Reducing Training Costs by Employment Testing." *Personnel Psychology* 9: 39-44; Spring 1956.
75. SEASHORE, CARL E., and OTHERS. *Seashore Measures of Musical Talents.* New York: Psychological Corporation, 1956.
76. SEGEL, DAVID. "The Multiple Aptitude Tests." *Personnel and Guidance Journal* 35: 424-32; March 1957.
77. SORENSON, GARTH, and SENIOR, NOEL. "Changes in GATB Scores with College Training." *California Journal of Educational Research* 6: 170-73; September 1955.
78. STALNAKER, JOHN M. "The Study of Applicants, 1954-1955." *Journal of Medical Education* 30: 625-36; November 1955.
79. SUPER, DONALD E. "The Multifactor Tests: Summing Up." *Personnel and Guidance Journal* 36: 17-20; September 1957.
80. SUPER, DONALD E. "The Use of Multifactor Test Batteries in Guidance." *Personnel and Guidance Journal* 35: 9-15; September 1956.
81. THORNDIKE, ROBERT L. "Development and Applications of Tests of Special Aptitudes." *Review of Educational Research* 26: 14-25; February 1956.
82. THURSTONE, LOUIS L. *Primary Mental Abilities.* Psychometric Monographs, No. 1. Chicago: University of Chicago Press, 1938. 121 p.
83. THURSTONE, THELMA G. "The Tests of Primary Mental Abilities." *Personnel and Guidance Journal* 35: 569-77; May 1957.
84. TRAXLER, ARTHUR E. "Relationship of Certain Predictive Measures to Achievement in First-Year French and Latin." *Educational Records Bulletin* 66: 73-77; July 1955.
85. VINEYARD, EDWIN E. "A Longitudinal Study of the Relationship of Differential Aptitude Test Scores with College Success." *Personnel and Guidance Journal* 36: 413-16; February 1958.
86. WEBB, SAM C. "The Prediction of Achievement for First Year Dental Students." *Educational and Psychological Measurement* 16: 543-48; Winter 1956.
87. WHITTINGTON, R. W. T. "The Assessment of Potential Musical Ability in Secondary School Children." *Journal of Educational Psychology* 48: 1-10; January 1957.
88. WING, HERBERT D. "The Measurement of Musical Aptitude." *Occupational Psychology* 31: 33-37; January 1957.
89. WORSENCROFT, ROBERT R. "The Effect of Training on the Spatial Visualizing Ability of Engineering Students." *Journal of Engineering Drawing* 19: 7-12; February 1955.

CHAPTER IV

Development and Applications of Tests of Educational Achievement

ROBERT L. EBEL and ROBERT E. HILL, JR.

THE MOST recent reviews of research in this area were prepared by Engelhart (31), and by Bloom and Heyns (11). Twenty-five years of research on educational testing were reviewed by Bayley and others (7).

Role of Testing in Education

The use of tests in education tends to increase both in quantity of testing done and in scope. Wood (107) traced the history of objective testing, emphasizing the values of such tests. Boag (13) reported that only about 7 cents per pupil per year was being used for standardized tests, which was far below the recommended minimum of 30 cents to 35 cents. Both amounts are insignificant compared with what a business may spend to determine the effectiveness of its practices.

Torbet (100) found generally unfavorable attitudes in secondary-school teachers toward teacher-made tests. Testing was seen as an onerous task. Substantial disagreement was found between expert recommendations and general practice in most aspects of test planning and construction.

Fricke and Millman (38) discussed the effects of high-school testing on students, faculty, parents, colleges, and employers, and reported many potential benefits. Wrightstone (110, 111) presented some basic ideas on the nature and interpretation of various tests, and discussed various ways in which different kinds of tests may be of benefit to students. Michael (69) discussed theoretical considerations and presented empirical findings related to differential testing of high-level personnel.

In large-scale testing programs, Tyler (102) reported data showing the effect of *General Educational Development* tests in the issuance of high-school equivalency certificates or diplomas for servicemen and veterans. *GED* accredited persons did as well as those possessing regular diplomas in the areas of industrial and public employment, and virtually as well in higher education. Barnette (6) discussed the role of college credit examinations at the University of Buffalo and traced the records of 205 individuals who had attempted credit by examination. Mallinson and Buck (67) reported favorably on various aspects of the New York State Regents Examinations in science.

Testing Techniques

Investigating the speed factor in testing, Lord (63) factorially analyzed both speeded and unspeeded tests of vocabulary, spatial relations, and

arithmetic reasoning, together with certain reference tests and academic grades. Four specific speed factors and a second-order general speed factor were isolated. Dole and Fletcher (26) suggested some principles for using incomplete sentences to measure (among other things) educational achievement. Thomas (96) investigated the use of construction-shift exercises in English as a measure of competence in written expression. He devised a procedure for objectively scoring these items. Morrissett (74) discussed the use of oral examinations at the college level. He found student opinion favorable to oral examinations and suggested that the total time involved in their use was not appreciably more than that required by written examinations. In an interesting experiment which related testing to learning, Gilbert (40) found that a procedure which enabled a student to know immediately upon choosing an answer whether it was correct did not appear to enhance learning nor did it facilitate test administration.

Essay Testing

Conflicting views were again expressed on the values of the essay test. Grant and Caplan (42) declared that short-answer essay examinations can be scored with quite satisfactory reliability and will discriminate adequately, provided special care is taken in construction and scoring. On the other hand, Pidgeon and Yates (82) reported on the reliability and validity of essay-type English papers written by 11-year-old children. Results indicated that even under ideal conditions, with a rigorous system of marking, essay-type papers do not achieve either the reliability or validity of objective tests.

Both students and teachers remain sympathetic to essay tests, at least if the latter are not used exclusively. In a controlled experiment with college students, Lundahl and Mason (65) found that weekly essay tests did not produce significantly greater gains in writing ability than weekly objective tests, but students favored the use of both types. French (37) reported a teacher questionnaire study of essay testing which was supported by the College Entrance Examination Board. A majority indicated that the *CEEB* examinations, which are largely objective, had no effect on their teaching. They strongly favored including an essay test in the program.

Test Development

Test development procedures were discussed both from the point of view of the professional expert and from that of the teacher. Tyler (103) and Lannholm (58) discussed the development of advanced-level tests in education. Tyler proposed an outline of specifications for such a test. Lannholm explained the merits of combining the work of the subject-matter specialist and the test expert. Similarly, Epstein and Myers (32) described the co-operation of teachers and test specialists in the production of a standardized test in mathematics. Zirkle and Austin (112)

reported on the co-operative efforts of a college faculty in developing a comprehensive social science examination. Dyer (28) discussed the rationale for the College Board's *Tests of Developed Abilities.* The tests were described by Coffman (28), who also explained the use of committees of expert teachers in their development. The growing practice of involving both subject-matter and test specialists in co-operative test construction efforts has many potentially valuable implications.

Objective Test Items

Flanagan (34) discussed items designed to measure important outcomes other than factual content, i.e., understandings, comprehensions, and applications. Several sample items were presented. Diederich, Nedelsky, and Engelhart (25) discussed the art of writing test exercises and presented examples of new, productive approaches. Bakan (4) compared results and opinions using multiple-choice items in the traditional way and also modified so that the examinee marked as many alternatives as needed to be "sure" of marking the correct answer. Student preference was about evenly divided for the two approaches, and performance was virtually the same in both cases. Moore (73) investigated five special response and scoring procedures for multiple-choice vocabulary test items, and found no important advantages over conventional "rights" scoring. Friedman and Fleishman (39) gained significant reliability by including a "don't know" category in a multiple-choice test of aural discrimination. Clark (18) reported positional preferences in five-alternative multiple-choice items to be very weak although time pressure caused some decreasing frequency in selection of the fifth-place alternative. Smith (92) investigated use of a two-alternative multiple-choice item form in a vocabulary test. He found it superior to three- and four-alternative forms. Bennett and Doppelt (9) found synonym and antonym vocabulary items superior to other types and also relatively easy to prepare.

Test Administration

Major considerations in the area of test administration seemed to center on the factors of time and student motivation in response to stress. At the college level, Barch (5) found that voluntary persistance was related to test achievement. Cook (19) found that a time announcement at the half-way point during a reading test at the college level significantly affected the proportion of items correct out of items attempted in the case of slow readers, but not for fast readers. Johnson (54) discussed a procedure for shortening tests without important loss of validity. Bennett and Doppelt (8) found vocabulary ability, as well as item difficulty, related to speed of response. The slowest quarter of the group studied worked at about the same rate with both easy and difficult materials.

Carrier (16) noted detrimental effects of stress when experimentally manipulated during course achievement examinations. These effects were

varied and were greater for females than males. The reduction of stress by encouragement between subtests was found by Sinick (89) to affect favorably the test performance of low-anxiety students but not medium- and high-anxiety students. Flanagan (35) suggested a motivational index based on measures of the proportion of papers showing a formal marking pattern and the proportion of papers having scores at or below chance expectations.

Anderson (3) investigated attitudes of college students toward designation of certain behaviors as cheating. Females expressed stricter attitudes than males, and students of education were stricter than certain other curricular groups. Canning (15) reported from a questionnaire study that an honor system reduced cheating at the college level.

Item Analysis

Although recent contributions in statistical methodology to item selection and item analysis are considered in detail in Chapter VIII, it seems that certain papers representing efforts at simplified item-analysis procedures embody the implied hope that teachers will employ such techniques in constructing their own examinations. Cuadra (21) presented a simplified form and technique for item analysis using the Hankes-type answer sheet. Findley (33) offered a logical and mathematical analysis in support of a simplified item-discrimination procedure that is also cited in Chapter VIII of this issue.

A nonconventional approach was investigated by Tomlinson and Schmid (99), who noted that selecting items which discriminate both ways in a two-way classification (i.e., in both of two aptitude or achievement areas) reduced verbal variance. It was suggested that this procedure may have ramifications for determining suppressor variables and be useful in building test batteries which would efficiently predict several criteria. Kropp (56) analyzed verbalized recordings of processes used in solving test items. Inferring process from response was found to be very hazardous, but the procedure was effective in revealing item ambiguities, hidden clues, and the like.

Test Analysis

Stanley (93) discussed simplified means of determining test analysis statistics. Rinsland (86) devised a standard check form for evaluating standardized tests. Lord (64) found empirical support for the contention that easier tests tend to yield negative skewness; harder tests, positive skewness. There was some indication that symmetrical distributions tend to be platykurtic. Adams (1) discussed formulas for analyzing various types of objective tests which had been scored and marked, then returned to students for review purposes, subsequently re-marked by students, and returned to instructors for evaluation after the "second-guessing." Medley (68) found that two tests which were identical in content but different

in item form (true-false versus multiple-choice) could not be regarded as equivalent with respect to their mean although their variance and reliabilities were comparable for the samples studied.

Validity and Reliability

Most constructors and many users of educational achievement tests are aware of the fundamental problems involved in the creation of valid tests and the provision of evidence on test validity. Hence, it is somewhat surprising and disappointing that there have been so few clarifying discussions of the complex problems associated with test validity. Only construct validity has received much attention. Chapter VII deals specifically with developments in this area.

The validity of many educational achievement tests depends on the adequacy of their sampling of some specified body of content. Lennon (59) pointed out some of the assumptions underlying the use of content validity. Huddleston (52) discussed test development on the basis of content validity, and Ebel (29) described ways of obtaining and reporting evidence of content validity. The most troublesome problem associated with studies of predictive or concurrent validity is that of obtaining adequate criterion measures. The use of performance tests, disguised as classroom exercises, to obtain criterion measures of special talent in creative writing was described by Wilson (106).

One avenue for studying the validity of a test, which has not been sufficiently explored, is that which investigates the correlation of scores from it with scores from other allegedly different tests. Wright and Scarborough (108) found relatively high correlations between the *Area Tests* of the *Graduate Record Examinations* and the *Cooperative General Culture Test*. This led them to question the difference between the two tests. The technical manual for the *SRA Achievement Series* (97) describes a factor analysis of scores on a battery including tests of reading, arithmetic, language, and study skills. The analysis produced a general achievement factor and specific factors for reading, language, and arithmetic. There was no specific factor for study skills. That is, when the study skills tests were administered to this group of students, they apparently did not measure anything which could be distinguished from the kind of general achievement measured by the other tests in the battery. This raises questions concerning the practical value of the study skills tests. It is important to note that the publishers of comparable batteries have generally not even investigated whether unique contributions were made by the several tests in their batteries.

Developments in the estimation of test reliability are covered in detail in Chapter VIII. Here it is only appropriate to mention studies having to do with the interpretation of reliability coefficients. The standard error of measurement, calculated from the reliability coefficient and the variance of test scores, has often been regarded as an alternative, and in some

respects superior, indication of test reliability. Lord (61) argued convincingly that the standard error of measurement for a test is essentially a function of its length and is not affected appreciably by other characteristics which affect the over-all quality of the test.

Norms and Score Interpretation

One of the most frequent criticisms of published standardized tests has been that the norms are inadequate. In recent years test publishers have been working co-operatively to improve the comparability of norms for similar tests. Lennon (60) described a method of achieving this goal which combined the use of related standardizing populations with the use of a reference test.

Norms for many tests are based on data whose availability is at least partly fortuitous. Lamke (57) pointed out that only through the use of statistically appropriate sampling techniques is it possible to make exact estimates of the size of sampling errors in the norms data. Hagen and Thorndike (46) described a project in which a house-to-house survey, including the administration of a test, was used to obtain normative data for adult males. An important comparative study of 1943 and 1955 norms for the *USAFI Tests of General Educational Development* was reported by Bloom (10) and by Bloom and Statler (12). High-school seniors tested during the last two months of the 1954-55 school year showed marked improvement over those tested in 1943 in a majority of states. The wide differences found among the states were highly related to differences in financial support for education, in the level of formal education of the adult population, and in the extent to which young people made use of existing educational facilities. Rhule (85) found that the test performance of military personnel on the *USAFI Subject Examinations* was quite similar in the main to that of the civilian standardization groups.

Largely for reasons of mathematical convenience many test publishers assume that achievement growth is linear throughout the 10 months of the school year. North (75), however, found that most of the achievement gains on the *Stanford Achievement Test* were registered during a six-month period between fall and spring testings. Traxler (101) investigated the hypothesis that a selected pupil population, such as that found in the independent schools tested by the Educational Records Bureau, would yield lower correlations among test scores because of the restricted range of ability. Actually, the distributions of scores and the coefficients derived from them were not markedly different from those obtained with public-school pupils.

Test publishers are frequently urged to provide more specialized norms for diverse groups of examinees. A steep curve of diminishing returns tends to limit both the significance and the applicability of norms for narrowly specific groups. Nevertheless, it is often useful to obtain and

compare norms for clearly defined special groups. Hiskey (49) reported norms for children with normal hearing on the *Nebraska Test of Learning Aptitude* which was originally standardized on children with impaired hearing. Otterness and others (79) reported trade-school norms for some commonly used tests. Osborne and Sanders (78) found that recency of training, age of subjects, and type of undergraduate training were important variables influencing performance on the *Graduate Record Examination*. They suggested specialized norming to allow for the influence of these variables.

The interpretation of achievement test scores has frequently been related to measures of intellectual or academic ability, with inferences of over- or under-achievement. Tiedeman and McArthur (98) challenged the logic underlying this conception. They concluded that most cases seeming to reflect over- or under-achievement could actually be accounted for by errors of measurement or by other sources of unpredicted variance. They reported that these sources of unpredicted variance seemed to have very little to do with educational, personality, or interpersonal problems of the students.

Quotients can be used to express level of development of characteristics other than intelligence. DeLong (23) found that height and arithmetic quotients were more stable than spelling, weight, or reading quotients, with intelligence quotients least stable of all. He questioned some of the assumptions involved in the use of quotients. He recommended that no score be considered as representing a characteristic value for a person unless three or more measures of it are available. He also suggested that no judgment about intellectual capacity be made until such a value has been estimated repeatedly over a period of time.

Many educators would support the idea that evaluations of educational achievement are more properly based on measures of growth than on measures of status. However, there is considerable theoretical and experimental evidence that measures of growth are often likely to be quite unreliable. In addition to the important analytic formulations of growth measurement proposed by Lord (62) and by McNemar (66), that are described in Chapter VIII, Diederich (24) called attention to data showing that on measures of growth, the least able students have a considerable advantage over the most able. He attributed this to regression effects, the use of tests too easy for the most able students, and inequality in the units of measurement.

Some test specialists contend that there is little difference between the qualities measured by the typical group intelligence test and the typical achievement test battery, at least at the elementary-school level. North (76) studied the relationship between *Kuhlmann-Anderson* IQ's and *Stanford Achievement Test* scores. While the relationship he found was substantial, he judged it not high enough to signify that the two tests were measuring the same abilities. He suggested that the apparent over-

or under-achievement of certain schools might reflect variations in the effectiveness of their instructional programs or might indicate differences in the closeness with which the content of the achievement test paralleled their curriculums.

Uses of Achievement Tests

There is general agreement that current deficiencies in educational measurement are more the result of inadquate or improper use of test results than of inadequate test instruments. (This can be said without implying that currently available tests are beyond serious criticism.) If the problem of better test utilization has not been solved, it is not because of lack of discussion of it. Seashore and Dobbin (88) offered suggestions for more effective use of test results.

Improvements in curriculums and in guidance as a result of special training of college professors in the use of test results were reported by Honora and Steible (51). Smith (90) criticized improper and inadequate college examinations and showed their adverse effects on college marks. Miller (71) and Eley (30) discussed testing in the language arts. Other discussions of improved use of test results were provided by Smith (91) and by Seashore (87). At a recent conference (72) Chauncey, Dressel, Coffman, and Mayhew described problems and developments in evaluation in higher education.

The educational impact and benefits of a rigorous, high-level scholarship selection examination procedure were pointed out by Holland and Stalnaker (50). Hacker (45), taking note of the growing trend toward adult education and irregular collegiate continuation education for young adults, recommended that urban universities take on an external degree-granting function on the basis of examinations. Peters (81) found evidence that the *USAFI GED* tests are effective measures of educational achievements acquired through nonacademic experiences. Hill (47) made use of the results of a wide-scale testing program to compare the high-school achievements of students who came from different elementary-school backgrounds—public urban, parochial, and rural. He found public urban pupils superior in general to both rural and parochial pupils at the ninth-grade level. These differences persisted through the twelfth grade even when the pupils from various elementary-school backgrounds attended the same public school. Wesman (105) pointed out that while it is the publisher's obligation to prepare sound, modern, practical test instruments, it is the user's obligation to co-operate in the development of such tests and to support those which deserve support by purchase and by intelligent application and interpretation.

Prediction of Academic Success

There was continuing interest in the prediction of academic success. Hill (48), Hyman (53), Knoell (55), and Patterson (80) discussed the

use of various tests in predicting academic success at various educational levels and in various areas of specialization. French (36) compared the validity of the College Board's *Scholastic Aptitude Test* with the validities of several short experimental tests involving achievement materials. Tests in government and literature information were the most successful among the experimental tests. Their estimated validity, if made as long as the *Scholastic Aptitude Test,* was somewhat higher than that observed for the *SAT.*

There was greatly increased interest in the identification and education of the academically talented. Chauncey (17) pointed out the value of aptitude tests in academic selection. Piekarz and others (83) suggested means for identifying superior learners from kindergarten to college. While the major emphasis was placed on objective tests, other bases of identification were also discussed.

The value of tests in sectioning students to improve instructional efficiency has been stressed more often than it has been studied. Gustad and Fish (43) found that a single test in English achievement served as an effective selector for exempting students from elementary course work in English at the college freshman level. They also found that students who elected to be exempt achieved more highly in subsequent work than students of similar ability who elected not to take advantage of the opportunity.

New Tests

Borg and Goodman (14) reported that while group tests of English for foreign students appeared to measure comprehension satisfactorily, they were less satisfactory as a measure of ability in English expression. A new individual oral test, consisting of 60 questions based on a simplified model of a flying training base, was shown to have satisfactory reliability and encouraging indications of validity. Wrightstone (109) described construction of tests of mathematical concepts for young children, outlining the content, development, standardization, and analysis of the tests. Orleans and Lindberg (77) cited evidence of serious deficiency in arithmetic understanding among teachers and pointed to the lack of a test specifically designed to measure this variable. The procedures they used in developing two 16-item forms of a test to measure arithmetic understanding were described. Habel (44) discussed a mathematics test appropriate at the college freshman level. The test was designed to have special advantages as a power test and as a diagnostic instrument.

Douglas (27) considered the relation of college admission tests to secondary-school curriculums in mathematics. The test maker's problem is to keep up with a changing curriculum without directing that change or being unfair to students taking the newer, or the more traditional, courses. Wagner (104) described a test of economic knowledge and attitudes used in a workshop on economic education for teachers. Aliferis and Stecklein (2) discussed a new test designed to diagnose and evaluate three areas of music achievement—melody, harmony, and rhythm.

Stecklein (94) reported that different instrumental groups made significantly different scores on these sections of the *Aliferis Music Achievement Test.*

Remmlein (84) described a process of constructing an objective test in school law. Five school-law instructors were used as item critics.

Tarasow (95) reported an experiment in standardizing a Hebrew achievement test for the second year. Glickman (41) discussed the development of a new *Naval Knowledge Test* designed to predict the better risks for officer training.

A test of group problem solving in which the task is to copy a model made with a construction-type toy was described by Damrin (22). Each examinee is given one or two of the blocks needed to build the model. The group is allowed to discuss the problem freely and to develop a plan of action. The group score is the time required to copy the model. No individual scores are obtained. Miller (70) developed a new test for reasoning ability which required the recognition of fallacies in reasoning. Crowell and Dole (20) reported a test of animistic thinking which showed a moderate relationship with intelligence in a sample of college students.

Conclusion

This survey of research on the development and application of tests of educational achievement from 1955 to 1958 reveals important progress. Test constructors and test users reveal considerable ingenuity and growing sophistication, but there is obviously room for much future progress toward better solutions of three basic problems: What needs to be measured? How can the measuring devices be made more accurate and efficient? How can the measures be used more effectively to contribute to the total educational effort? As more precise analyses of these problems are made and as more adequate experimental procedures are brought to bear on them, more rapid progress can be anticipated.

Bibliography

1. ADAMS, SIDNEY. "Analysis of 'Second-Guessed' Training Tests: Improvement on Objective Choice Tests Which Are Reviewed and Re-Marked After Initial Correction." *Journal of Educational Research* 50: 533-42; March 1957.
2. ALIFERIS, JAMES, and STECKLEIN, JOHN E. "Measurement of Music Achievement at College Entrance." *Journal of Applied Psychology* 39: 263-72; August 1955.
3. ANDERSON, WILLIAM F., JR. "Attitudes of University Students Toward Cheating." *Journal of Educational Research* 50: 581-88; April 1957.
4. BAKAN, RITA. "The Use of a Modified Multiple Choice Item Under Various Conditions." *Journal of Educational Research* 51: 223-28; November 1957.
5. BARCH, ABRAM M. "The Relation of Departure Time and Retention to Academic Achievement." *Journal of Educational Psychology* 48: 352-58; October 1957.
6. BARNETTE, W. LESLIE, JR. "Advanced Credit for the Superior High School Student." *Journal of Higher Education* 28: 15-20; January 1957.
7. BAYLEY, NANCY, and OTHERS. "Educational Measurement." *Review of Educational Research* 26: 268-91; June 1956.

8. BENNETT, GEORGE K., and DOPPELT, JEROME E. "Item Difficulty and Speed of Response." *Educational and Psychological Measurement* 16: 494-96; Winter 1956.
9. BENNETT, GEORGE K., and DOPPELT, JEROME E. "Relative Efficiency of Seven Verbal Item Types." *Educational and Psychological Measurement* 16: 497-500; Winter 1956.
10. BLOOM, BENJAMIN S. "1955 Normative Study of the Tests of General Educational Development." *School Review* 64: 110-24; March 1956.
11. BLOOM, BENJAMIN S., and HEYNS, I. DE V. "Development and Applications of Tests of Educational Achievement." *Review of Educational Research* 26: 72-88; February 1956.
12. BLOOM, BENJAMIN S., and STATLER, CHARLES R. "Changes in the States on the Tests of General Educational Development from 1943 to 1955." *School Review* 65: 204-21; June 1957.
13. BOAG, AUDREY K. "Standardized Tests: How, When, Why." *Instructor* 65: 24; October 1955.
14. BORG, WALTER R., and GOODMAN, JOHN S. "Development of an Individual Test of English for Foreign Students." *Air Force Human Engineering, Personnel, and Training Research, Technical Report* 56-8: 25-30; 1956.
15. CANNING, RAY R. "Does an Honor System Reduce Classroom Cheating? An Experimental Answer." *Journal of Experimental Education* 24: 291-96; June 1956.
16. CARRIER, NEIL A. "The Relationship of Certain Personality Measures to Examination Performance Under Stress." *Journal of Educational Psychology* 48: 510-20; December 1957.
17. CHAUNCEY, HENRY. "How Tests Help Us Identify the Academically Talented." *NEA Journal* 47: 230-31; April 1958.
18. CLARK, EDWARD L. "General Response Pattern to Five-Choice Items." *Journal of Educational Psychology* 47: 110-17; February 1956.
19. COOK, DESMOND L. "A Comparison of Reading Comprehension Scores Obtained Before and After a Time Announcement." *Journal of Educational Psychology* 48: 440-46; November 1957.
20. CROWELL, DAVID H., and DOLE, ARTHUR A. "Animism and College Students." *Journal of Educational Research* 50: 391-95; January 1957.
21. CUADRA, CARLOS A. "A New Technique for Rapid Item Analysis." *Journal of Applied Psychology* 40: 187-88; June 1956.
22. DAMRIN, DORA E. "The Russell Sage Social Relations Test: A Measure of Group Problem-Solving Skills in Elementary School Children." *12th Yearbook, National Council on Measurements Used in Education.* (Part 2) New York: the Council (Secy.-Treas.: Robert D. North, 21 Audubon Avenue), 1955. p. 5.
23. DELONG, ARTHUR R. "The Meaning of Individual Scores on Group Tests." *14th Yearbook, National Council on Measurements Used in Education.* New York: the Council (Secy.-Treas.: Robert D. North, 21 Audubon Avenue), 1957. p. 43-49.
24. DIEDERICH, PAUL B. "Pitfalls in the Measurement of Gains in Achievement." *School Review* 64: 59-63; February 1956.
25. DIEDERICH, PAUL B.; NEDELSKY, LEO; and ENGELHART, MAX D. "Exercise Writing in the Humanities, Natural Sciences, and Social Sciences." *Proceedings of the 1957 Individual Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1958. p. 36-37.
26. DOLE, ARTHUR A., and FLETCHER, FRANK M., JR. "Some Principles in the Construction of Incomplete Sentences." *Educational and Psychological Measurement* 15: 101-10; Summer 1955.
27. DOUGLAS, EDWIN C. "College Board Examinations and Curriculum Change." *Mathematics Teacher* 50: 305-308; April 1957.
28. DYER, HENRY S., and COFFMAN, WILLIAM E. "The Tests of Developed Abilities." (Symposium) *College Board Review* No. 31: 5-10; Winter 1957.
29. EBEL, ROBERT L. "Obtaining and Reporting Evidence on Content Validity." *Educational and Psychological Measurement* 16: 269-82; Autumn 1956.
30. ELEY, EARLE G. "Testing the Language Arts." *Modern Language Journal* 40: 310-15; October 1956.
31. ENGELHART, MAX D. "Testing and Use of Test Results." *Review of Educational Research* 26: 5-13; February 1956.

32. EPSTEIN, MARION, and MYERS, SHELDON S. "How a Mathematics Test Is Born." *Mathematics Teacher* 51: 299-302; April 1958.
33. FINDLEY, WARREN G. "A Rationale for Evaluation of Item Discrimination Statistics." *Educational and Psychological Measurement* 16: 175-80; Summer 1956.
34. FLANAGAN, JOHN C. "Can We Measure What We Teach?" *High School Journal* 41: 93-96; December 1957.
35. FLANAGAN, JOHN C. "The Development of an Index of Examinee Motivation." *Educational and Psychological Measurement* 15: 144-51; Summer 1955.
36. FRENCH, JOHN W. "Validation of New Item Types Against Four-Year Academic Criteria." *Journal of Educational Psychology* 49: 67-76; April 1958.
37. FRENCH, JOHN W. "What English Teachers Think of Essay Testing." *English Journal* 46: 196-201; April 1957.
38. FRICKE, BENNO G., and MILLMAN, JASON. "Who Benefits from High School Testing?" *High School Journal* 41: 71-74; December 1957.
39. FRIEDMAN, MORTON, and FLEISHMAN, EDWIN A. "A Note on the Use of a 'Don't Know' Alternative in Multiple Choice Tests." *Journal of Educational Psychology* 47: 344-49; October 1956.
40. GILBERT, ARTHUR C. F. "Effect of Immediacy of Knowledge of Correctness of Response upon Learning." *Journal of Educational Psychology* 47: 415-23; November 1956.
41. GLICKMAN, ALBERT S. "The Naval Knowledge Test." *Journal of Applied Psychology* 40: 389-92; December 1956.
42. GRANT, DONALD L., and CAPLAN, NATHAN. "Studies in the Reliability of the Short-Answer Essay Examination." *Journal of Educational Research* 51: 109-16; October 1957.
43. GUSTAD, JOHN W., and FISH, JANICE P. "The Use of the Cooperative Mechanics of Expression Test in Classification at the College Freshman Level." *Educational and Psychological Measurement* 15: 436-40; Winter 1955.
44. HABEL, ELMER A. "Implications Arising out of Students' Errors." *Journal of Higher Education* 29: 81-88; February 1958.
45. HACKER, LOUIS M. "New Kinds of Students and New Ways of Testing Achievement." *Proceedings of the 1956 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1957. p. 95-102.
46. HAGEN, ELIZABETH P., and THORNDIKE, ROBERT L. "Normative Test Data for Adult Males Obtained by House to House Testing." *Journal of Educational Psychology* 46: 207-16; April 1955.
47. HILL, ROBERT E., JR. "An Investigation of the Educational Development of Selected Iowa Secondary School Pupils from Varied Elementary School Environments." *14th Yearbook, National Council on Measurements Used in Education.* New York: the Council (Secy.-Treas.: Robert D. North, 21 Audubon Avenue), 1957. p. 28-36.
48. HILL, SUZANNE D. "The Relationship Between Grades and a Predictive Test Battery in the School of Pharmacy of the George Washington University." *Journal of Applied Psychology* 41: 61-62; February 1957.
49. HISKEY, MARSHALL S. "Norms for Children with Hearing for the Nebraska Test of Learning Aptitude." *Journal of Educational Research* 51: 137-42; October 1957.
50. HOLLAND, JOHN L., and STALNAKER, JOHN M. "An Honorary Scholastic Award." *Journal of Higher Education* 28: 361-68; October 1957.
51. HONORA, SISTER MARY, and STEIBLE, DANIEL J. "A College Examines Its Use of Test Results." *Journal of Educational Research* 50: 611-15; April 1957.
52. HUDDLESTON, EDITH M. "Test Development on the Basis of Content Validity." *Educational and Psychological Measurement* 16: 283-93; Autumn 1956.
53. HYMAN, SIDNEY R. "The Miller Analogies Test and the University of Pittsburgh Ph.D.'s in Psychology." *American Psychologist* 12: 35-36; January 1957.
54. JOHNSON, A. PEMBERTON. "The Development of Shorter and More Useful Selection Tests." *Journal of Educational Psychology* 46: 402-407; November 1955.
55. KNOELL, DOROTHY M. "A Second Attempt To Predict Teaching Success from Word Fluency Data." *Journal of Educational Research* 49: 13-25; September 1955.
56. KROPP, RUSSELL P. "The Relationship Between Process and Correct Item Response." *Journal of Educational Research* 49: 385-88; January 1956.

57. LAMKE, TOM A. "The Standardization of the Henmon-Nelson Revision." *13th Yearbook, National Council on Measurements Used in Education.* New York: the Council (Secy.-Treas.: Robert D. North, 21 Audubon Avenue), 1956. p. 42-44.
58. LANNHOLM, GERALD V. "The Development of an Advanced Level Test in Education." *Journal of Educational Research* 49: 311-13; December 1955.
59. LENNON, ROGER T. "Assumptions Underlying the Use of Content Validity." *Educational and Psychological Measurement* 16: 294-304; Autumn 1956.
60. LENNON, ROGER T. "Efforts Toward Greater Comparability of Norm Groups." *14th Yearbook, National Council on Measurements Used in Education.* New York: the Council (Secy.-Treas.: Robert D. North, 21 Audubon Avenue), 1957. p. 128-30.
61. LORD, FREDERIC M. "Do Tests of the Same Length Have the Same Standard Errors of Measurement?" *Educational and Psychological Measurement* 17: 510-21; Winter 1957.
62. LORD, FREDERIC M. "The Measurement of Growth." *Educational and Psychological Measurement* 16: 421-37; Winter 1956.
63. LORD, FREDERIC M. "A Study of Speed Factors in Tests and Academic Grades." *Psychometrika* 21: 31-50; March 1956.
64. LORD, FREDERIC M. "A Survey of Observed Test-Score Distributions with Respect to Skewness and Kurtosis." *Educational and Psychological Measurement* 15: 383-89; Winter 1955.
65. LUNDAHL, WALTER S., and MASON, JOHN M. "Essay Testing in Biological Science as a Means for Supplementing Training in Writing Skills." *Science Education* 40: 261-67; October 1956.
66. MCNEMAR, QUINN. "On Growth Measurement." *Educational and Psychological Measurement* 18: 47-55; Spring 1958.
67. MALLINSON, GEORGE G., and BUCK, JACQUELINE V. "An Investigation of the New York State Regents Examinations in Science." *Journal of Experimental Education* 24: 43-89; September 1955.
68. MEDLEY, DONALD M. "The Influence of Item Modality on the Dimension Measured by a Test." *Journal of Experimental Education* 24: 303-307; June 1956.
69. MICHAEL, WILLIAM B. "Differential Testing of High Level Personnel." *Educational and Psychological Measurement* 17: 475-90; Winter 1957.
70. MILLER, ELMER H. "A Study of Difficulty Levels of Selected Types of Fallacies in Reasoning and Their Relationships to the Factors of Sex, Grade Level, Mental Age, and Scholastic Standing." *Journal of Educational Research* 49: 123-29; October 1955.
71. MILLER, PETER M. "How Much Testing and What Kinds of Tests in the English Language Arts?" *Bulletin of the National Association of Secondary-School Principals* 39: 91-98; September 1955.
72. MILLER, ROBERT D., editor. *Program and Proceedings of the Conference on Evaluation in Higher Education.* Tallahassee: Florida State University, 1954. 88 p.
73. MOORE, ROBERT. "A Comparison of Selected Modifications of a Multiple Choice Examination." Doctor's thesis. Iowa City: State University of Iowa, 1956. 150 p. Abstract: *Dissertation Abstracts* 16: 1844; No. 10, 1956.
74. MORRISSETT, IRVING. "An Experiment with Oral Examinations." *Journal of Higher Education* 29: 185-90; April 1958.
75. NORTH, ROBERT D. "Achievement Growth Trends of Independent School Pupils as Reflected by Fall and Spring Results on the Stanford Achievement Test." *Educational Records Bulletin,* No. 66, July 1955, p. 57-68.
76. NORTH, ROBERT D. "Relationship of Kuhlmann-Anderson IQ's and Stanford Achievement Test Scores of Independent School Pupils." *Educational Records Bulletin,* No. 68, July 1956, p. 53-60.
77. ORLEANS, JACOB S., and LINDBERG, LUCILE. "The Preparation of a Test of Arithmetic Understanding." *12th Yearbook, National Council on Measurements Used in Education.* (Part 2) New York: the Council (Secy.-Treas.: Robert D. North, 21 Audubon Avenue), 1955. p. 51-56.
78. OSBORNE, R. TRAVIS, and SANDERS, WILMA B. "Recency and Type of Undergraduate Training and Decline of G.R.E. Performance with Age." *Journal of Educational Psychology* 47: 276-84; May 1956.

79. OTTERNESS, WILLIAM B., and OTHERS. "Trade School Norms for Some Commonly Used Tests." *Journal of Applied Psychology* 40: 57-60; February 1956.
80. PATTERSON, CECIL H. "Predicting Success in Trade and Vocational School Courses: Review of the Literature." *Educational and Psychological Measurement* 16: 352-400; Autumn 1956.
81. PETERS, FRANK R. "Measurement of Informal Educational Achievement by the GED Tests." *School Review* 64: 227-32; May 1956.
82. PIDGEON, D. A., and YATES, ALFRED. "Symposium: The Use of Essays in Selection at 11+: IV. Experimental Inquiries into the Use of Essay-Type English Papers." *British Journal of Educational Psychology* 27: 37-47; February 1957.
83. PIEKARZ, JOSEPHINE A., and OTHERS. "Identification of Superior Learners." *Supplementary Educational Monographs*, No. 81. Chicago: University of Chicago Press, 1954. Chapter 3, p. 22-34.
84. REMMLEIN, MADALINE K. "The Construction of an Objective Test in School Law—A Preliminary Report." *13th Yearbook, National Council on Measurements Used in Education*. New York: the Council (Secy.-Treas.: Robert D. North, 21 Audubon Avenue), 1956. p. 10-23.
85. RHULE, WARREN A. "The Performance of Military Personnel on USAFI Subject Examinations." *Journal of Educational Research* 51: 541-44; March 1958.
86. RINSLAND, HENRY D. "A Form for Briefing and Evaluating Standardized Tests." *Journal of Educational Research* 49: 629-32; April 1956.
87. SEASHORE, HAROLD G. "Tests as Aids to Administration and Counseling in Junior Colleges." *Junior College Journal* 26: 504-508; May 1956.
88. SEASHORE, HAROLD G., and DOBBIN, JOHN E. "How Can the Results of a Testing Program Be Used Most Effectively?" *Bulletin of the National Association of Secondary-School Principals* 42: 64-68; April 1958.
89. SINICK, DANIEL. "Encouragement, Anxiety, and Test Performance." *Journal of Applied Psychology* 40: 315-18; October 1956.
90. SMITH, C. PAGE. "Human Time and the College Student." *Journal of Higher Education* 28: 70-74; February 1957.
91. SMITH, EUGENIA. "Testing Can Be Teaching." *Practical Home Economics* 34: 10-11; December 1955.
92. SMITH, KENDON. "An Investigation of the Use of 'Double-Choice' Items in Testing Achievement." *Journal of Educational Research* 51: 387-89; January 1958.
93. STANLEY, JULIAN C., JR. "Simplified Test Analysis Statistics." *Journal of Higher Education* 27: 498-500; December 1956.
94. STECKLEIN, JOHN E. "Relationship of Instrument to Music Achievement Test Scores." *13th Yearbook, National Council on Measurements Used in Education*. New York: the Council (Secy.-Treas.: Robert D. North, 21 Audubon Avenue), 1956. p. 146-50.
95. TARASOW, MORRIS. "An Experiment in Standardizing a Hebrew Achievement Test for the Second Year." *Jewish Education* 26: 51-55; No. 3, 1956.
96. THOMAS, MACKLIN. "Construction Shift Exercises in Objective Form." *Educational and Psychological Measurement* 16: 181-86; Summer 1956.
97. THORPE, LOUIS P.; LEFEVER, D. WELTY; and NASLUND, ROBERT A. *Technical Supplement, SRA Achievement Series*. Second edition. Chicago: Science Research Associates, 1957. p. 20-29.
98. TIEDEMAN, DAVID V., and MCARTHUR, CHARLES C. "Over and Under Achievement: If Any." *13th Yearbook, National Council on Measurements Used in Education*. New York: the Council (Secy.-Treas.: Robert D. North, 21 Audubon (Avenue), 1956. p. 135-45.
99. TOMLINSON, HELEN, and SCHMID, JOHN, JR. "Use of a Difference-Score Criterion in Item Analysis." *Journal of Educational Research* 50: 373-81; January 1957.
100. TORBET, DAVID P. "The Attitude of a Select Group of Colorado Secondary School Teachers Toward Informal Teacher-Made Tests as Measured by a Projective Interview." *Journal of Educational Research* 50: 691-700; May 1957.
101. TRAXLER, ARTHUR E. "A Note on the Variability of the Independent School Population and on the Correlation of Test Scores for This Population as Compared with That for Public Schools." *Educational Records Bulletin*, No. 68, July 1956, p. 69-76.
102. TYLER, HARRY E. "GED Tests: Friends or Foes?" *California Journal of Secondary Education* 31: 66-71; February 1956.

103. TYLER, LOUISE L. "Brief Notes on Specifications for an Advanced Level Test in Education." *Journal of Educational Research* 51: 383-86; January 1958.
104. WAGNER, LEWIS E. "Testing Economic Knowledge and Attitudes." *Bulletin of the National Association of Secondary-School Principals* 40: 120-32; May 1956.
105. WESMAN, ALEXANDER G. "The Obligations of the Test User." *Proceedings of the 1955 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1956. p. 60-65.
106. WILSON, ROBERT C. "Improving Criteria for Complex Mental Processes." *Proceedings of the 1957 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1958. p. 13-20.
107. WOOD, BEN D. "Testing; Then and Now." *Proceedings of the 1956 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1957. p. 58-66.
108. WRIGHT, JOHN C., and SCARBOROUGH, BARRON B. "The Interrelationship of Area Test Scores and Cooperative General Culture Test Scores." *Journal of Educational Psychology* 48: 460-63; November 1957.
109. WRIGHTSTONE, J. WAYNE. "Constructing Tests of Mathematical Concepts for Young Children." *13th Yearbook, National Council on Measurements Used in Education.* New York: the Council (Secy.-Treas.: Robert D. North, 21 Audubon Avenue), 1956. p. 107-10.
110. WRIGHTSTONE, J. WAYNE. "Do Students Benefit from Testing?" *High School Journal* 41: 75-78; December 1957.
111. WRIGHTSTONE, J. WAYNE. "Tests and What They Test." *NEA Journal* 47: 221-22; April 1958.
112. ZIRKLE, GEORGE A., and AUSTIN, RONALD L. "Social Science Comprehensive Examination." *Journal of Higher Education* 27: 38-40; January 1956.

CHAPTER V

Development and Applications of Structured Tests of Personality

WILLIAM COLEMAN and DOROTHY MANLEY COLLETT

STRUCTURED tests of personality are broadly conceived in this chapter as including inventories of interests and values as well as those of personality and adjustment. Structured instruments require the respondent to choose one of a set of alternatives rather than requiring a quasi-idiosyncratic response to an ambiguous situation.

Since the 1956 review by Furst and Fricke (61), several new structured instruments appeared and a plethora of studies was made with some of the older inventories and a few of the newer ones.

The impact of work in the area of structured personality tests during the last three years may be generally evaluated as having contributed very little in the way of original concepts. The reported research supplies more empirical data for some of the instruments and delineates areas of fruitful use for these inventories with evidence frequently of concurrent validity and occasionally of predictive validity.

With the notable exception of the book by Stern, Stein, and Bloom (121), little has been done to advance work toward more carefully defined criteria. Thus, efforts at determining the validity of the various inventories are vitiated by the lack of clean-cut criteria. General contributions to personality measurement were made by Loevinger (93), Cronbach and Gleser (39), David and von Bracken (41), and Hall and Lindzey (71). Methodological advances in test construction as well as developments in pattern and profile analysis are discussed in Chapter VIII.

Since the general "state of the art" is discussed in Chapters VII and VIII, this chapter was organized to present a critical summary of the research data that were published relative to specific inventories. Although it is recognized that well-constructed instruments do not always receive attention in research studies, research is necessary to provide useful data enabling users to know more about the reliability and validity of an instrument.

The authors of this chapter made a diligent effort to search the literature and received assistance from some of the test publishers in locating published research studies that provide some evidence as to the validity or reliability of structured inventories. For most of the inventories only one or two or no research studies were found; whereas, the *Minnesota Multiphasic Psychological Inventory (MMPI)*, *Edwards Personal-Preference Schedule (EPPS)*, *Minnesota Teacher Attitude Inventory (MTAI)*, and the *Taylor Manifest Anxiety Scale (TMAS)* had 10 or more. Some of

the other inventories may have been the subject of much more research during the last three years than reported here, possibly as theses or dissertations. However, the user of most of these inventories must recognize that the general literature lacks research data which might enable him to evaluate the potential usefulness of an instrument for a given purpose.

Estimates of Validity

Twelve years ago, Ellis (49) reviewed the status of research on the validity of personality inventories, concluding that possibly only the *MMPI* as an individually administered scale might have some validity. More recently, Fiedler and others (54) and Tindall (133) examined the interrelationships of various indexes of adjustment. In both instances, the studies reported very little relationship among the instruments, even in instances when the same titles were used for subscales. Counselors and educational workers using adjustment inventories need to be aware of this lack of agreement as to what is being measured before glib interpretations are made of scores derived from such instruments.

Validity studies made during the last three years with a large number of the adjustment inventories also failed to show any empirical basis for supporting their validity. These studies usually involved the use of criterion groups based on peer ratings or ratings made by observers who were professional workers. Sometimes one scale was used for determining the criterion groups even though the evidence as to the validity of the scale used for that purpose was ambiguous. Scores of respondents in the criterion groups (usually college students) were then compared by use of a *t* test, or *F* when a slightly more sophisticated design was used. Since ratings are often unreliable and distinct categorizations of behavior or personality are not easily achieved, it is likely that the criterion categories used in most of these validation studies were contaminated. There are several examples of these studies (4, 9, 13, 14, 44, 79, 104, 120, 128, 132, 136).

Recognizing the difficulties inherent in attempting to establish clear-cut criterion groups based on rater judgment, Thurstone and various other psychometricians resorted to factor analysis to establish pure traits. Thus, Comrey (31, 32, 33, 34, 35, 36, 37) and Comrey and Marggraff (38) published the results of a series of factor-analytic studies with the *MMPI*, and Guilford and Zimmerman (70) described a factor study of three inventories developed by Guilford. Cattell (27) reviewed his factor-analytic work with the *Sixteen Personality Factor Questionnaire*, and Heron (76) factor-analyzed scores on 19 indexes for males and females separately.

Stewart (123) used four respondents to set up matrices with 27 personality inventory items as the variables in an interesting reversal of the usual factor analysis procedure. His nonhospitalized subjects showed a

greater concentration of common factor variance than his mental patients. This was interpreted as showing greater personality integration for the nonhospital group.

Fakability and Response Set

Although it has been well recognized in the literature that personality and interest inventories are usually susceptible to faking, contrary claims are often made by the authors of inventories. In this section, studies of the susceptibility to faking of some of the new inventories will be reviewed.

Sundberg and Bachelis (125) demonstrated that college students were able to fake prejudiced or unprejudiced scores on the *California F Scale* and the *PR Scale* of the *California Psychological Inventory*. Davids (42) demonstrated that the *Taylor Anxiety Scale* also was susceptible to deception. Studies by Mitzel and others (102), Della Piana and Gage (45), Stein and Hardy (120), and Sorenson (117) all demonstrated that the *MTAI* is appreciably susceptible to faking. However, in a further study, Sorenson and Sheldon (118) found that groups of respondents were not likely to fake unless they received a cue from the instructions. Della Piana and Gage (45) were also able to demonstrate that pupil values were significantly related to responses to the *MTAI*.

Three fakability studies (10, 67, 108) with the *Gordon Personal Profile*, a four-factor, forced-choice inventory, indicated that this instrument was only slightly susceptible to distortion. Students were instructed to fake their responses in a simulated industrial situation and a simulated guidance situation in Rusmore's study (108).

Borislow (18) investigated the fakability of the *EPPS* with respect to both personally and socially desirable items. He found it susceptible to faking, but the socially desirable attitude was less susceptible to distortion than the personally desirable set. Kaess and Witryol (82) determined that the *Pensacola Z Scale*, another forced-choice inventory, was only partially susceptible to distortion. In a well-designed study, Izard and Rosenberg (81) concluded that the *PRB Forced-Choice Leadership Test* was not easily susceptible to faking. Gehman (64) found the *Strong Vocational Interest Blank* highly susceptible to faking, a finding consonant with earlier studies of the *Strong*. Using items drawn from several personality inventories, Heron (75) demonstrated that response distributions under selection conditions were significantly different from response distributions under research conditions.

It would seem evident from such studies as cited above that forced-choice scales have substantially reduced the extent to which responses to inventories may be faked, thus enhancing the potential validity of inventories as a means of assessing personality. However, French (58) showed that use of a forced-choice scale with a population different from the standardization group requires revising the scale to assure that the alter-

natives remain equally desirable. A second method of coping with faking is to include a lie or faking score as is done on the *MMPI* and the *Kuder Preference Record.* Voas (135) has obtained some promising data by using a third procedure that requires respondents first to answer inventory questions as required and then to mark the most socially acceptable answer on a second answer sheet.

Related to the problem of fakability is the question of response set. Shelley's analysis (111) of investigations using the *California Attitude Scales* led him to conclude that possibly spurious reliability values and reduced validity may be attributed to the response set of "acquiescence." Chapman and Campbell (28) pointed out that acquiescence response set is an important factor in the *F Scale.* Their study involved reversing the wording of some of the items in order to study the effect of "agree" and "disagree" items on item reliability as well as response set. For two studies, the *F plus* was more reliable than the *F minus Scale.* Fricke (60) examined response set as a suppressor variable in the *OAIS* and *MMPI,* and suggested a method for constructing a scale to measure a testee's set to say "true." In the construction and use of personality tests the direction of the scored responses should be considered as well as the degree to which an item discriminates.

Mitzel and others (102) identified response sets of positive and negative intensity and of evasiveness in studying validity effects on the *MTAI.* The negative inventory set significantly increased test validity whereas the positive intensity set did not. Evasiveness was an attenuating influence on test validity. The response set theory derived by these authors suggests that correlations between attitude measures may be a function of a common response set instead of resulting from underlying relationships.

Development of Anxiety Scales

In the last few years, there has been considerable activity in the development and study of anxiety scales. The *Taylor Manifest Anxiety Scale (TMAS),* with items drawn from the *MMPI* pool and its derivatives (14), and the children's form (25) were the center of much research effort (21, 42, 43, 129, 131). These studies generally indicated that anxiety may be measured effectively through inventories. Davids and Eriksen (43) demonstrated a significant relationship between manifest anxiety as measured by the *TMAS* and productivity on a chained word association test. Taylor (130) reviewed the experimental studies concerned with drive theory and manifest anxiety, emphasizing that the *TMAS* was developed as a means of establishing anxiety level in order to investigate this variable in relation to drive theory. Although Taylor (131) found a relationship between *TMAS* scores and paired word association learning, Farber and Spence (53) questioned the adequacy of the evidence for a relationship between anxiety as a drive construct in learning theory and the effectiveness of learning.

A number of investigators (22, 43, 98, 109) examined the relationship of the *TMAS* to intelligence. Although studies with Air Force trainees (22) yielded significant negative correlations, studies using college students (92, 98) generally failed to obtain significant correlations, suggesting that specific testing conditions or the population used were major determinants of the extent of relationship between the *TMAS* and intelligence measures.

Item analysis studies (14, 21) of the *TMAS* suggested that a further reduction in the length of the scale (50 items compared with the 550 in the parent *MMPI*) might be profitable. Heineman's suggested item format was employed by Christie and Budnitzky (29) for the 20 items that Bendig (14) had shown to have clinical validity. Reliability data are reported for this forced-choice form with 20 items, but validity is assumed from previous studies.

Dreger and Aiken (46) derived a *Number Anxiety Scale* for the *TMAS* correlating .33 with the *Taylor Scale.* Number anxiety did not seem to be correlated to measures of intelligence, but was significantly correlated (—.44 and —.55 for two samples) with college mathematics grades.

The *Children's Form* of the *Manifest Anxiety Scale* was developed (25, 26, 95, 139), and studies were made with it, relating anxiety to school learning, complex learning tasks, and clinical anxiety.

To determine the validity of the *Sarason Test Anxiety Scale,* Martin and McGowan (97) used measures of palmar skin conductance as a criterion. The high-anxiety group on this *Anxiety Scale* had significantly higher skin conductance suggesting that Sarason's *Scale* might also be measuring a general anxiety factor.

Sinick (115) compared mean male and female scores on the *Sarason Test Anxiety Scale* and the *TMAS* and computed correlations. Females made higher mean scores on both instruments and had greater variances. The *r* between the scales for the 211 college students in Sinick's study was .43.

Relationship Between Personality and Vocational Interest

Several studies were made in the three years covered by this report on the relationship between personality and vocational interest. Melton (100) obtained 24 *r*'s significant at the 10- or 5-percent level between components of the *California Test of Personality* and *Vocational Interest Analysis.* Goodling (65) compared *Interest Maturity* scores on the *Strong Vocational Interest Blank (SVIB)* and the 10 trait scales of the *Guilford-Zimmerman.* With an N of 239, 4 *r*'s were significant at the 1-percent level with the highest *r* being .32. Klugman (86) compared the *Kuder Preference Record* scores of psychotic and neurotic veterans with the norm groups. After reviewing 30 studies, Patterson (103) concluded that emo-

tionally disturbed people tend to be more frequently interested in talent occupations or in social service type of work.

MMPI Research

The *MMPI* remained the most popular inventory for research studies during the period covered by this review. It was used to predict teaching success and general college achievement, to measure work attitude, to select medical students as well as graduate students in public health, and to relate personality variables to roles of union business agents.

Although Gowan and Gowan (68) found substantial correlations (.71, .75, and .83, corrected, for three groups) between *MMPI* scores and ratings of teaching candidates, LaBue (90) was unable to find a single *MMPI* score that correlated significantly with persistence in teaching. The *MMPI* added very little to the multiple *R* for predicting college average when Frick and Keener (59) ran a cross validation of a previous study.

Dominance and work attitude scales drawn from the *MMPI* items differentiated significantly between "good" and "poor" work attitude groups (132). The *MMPI* did not add to a multiple *R* for selecting graduate students in public health education (8), but the *L Scale* did contribute in selecting successful medical students (77). The size of a sample of union business agents was too small to permit the drawing of any conclusions from an investigation by Rosen and Rosen (107).

Peek and Storms (104) provided validity data for the Marsh-Hilliard-Liechti *MMPI Sexual Deviation Scale*, while cross-cultural comparisons with the *MMPI* were made by Taft (127). The social desirability of items in various *MMPI* scales was studied by Fordyce (56) and Hanley (72).

Barnes (5) obtained empirical data which seemed to support the hypothesis that atypical answers are associated with a "psychotic factor," and atypical false answers seem to be related to a neurotic factor. The data seemed to be consonant with Berg's "deviation hypothesis" (6).

Calvin and Hanley (23) used the Keeler polygraph on 13 subjects who had "faked good" and four who had "faked bad," all of whom had been selected from an original group of 300. Control groups were also established, and when comparisons were made, no significant differences were obtained between the faking groups and the controls.

To distinguish depressives from nondepressive psychotics, 26 face-valid items of the 60 in the *D Scale* were used to discriminate between the two groups as well as the entire scale. Winter and Salcines (138) showed that the Peterson *MMPI Psychosis Scale* is effective in predicting whether a person is psychotic.

Matarazzo (98) demonstrated a substantial correlation between scores on the *MTAI* and the *L*, *F*, and *K* validity scales of the *MMPI*. However, the overlap between the criterion measure items and the items on the *TMAS* may partially account for the high *r*'s which were obtained.

For people interested in examining the research with the *MMPI* in greater detail, Welsh and Dahlstrom (137) published a book containing 66 selected articles and nearly 700 references on the *MMPI*.

Minnesota Teacher Attitude Inventory (MTAI)

The *MTAI* continued during the last three years to be used frequently in research studies. In the June 1958 issue of the REVIEW devoted to teacher personnel (7), 14 studies with the *MTAI* were cited. These will not be reviewed again here.

Leeds (91) computed *r*'s between 10 *Guilford-Zimmerman Temperament Survey* scores and the *MTAI*, obtaining seven significant *r*'s. Traits most closely related to *MTAI* scores were personal relations, friendliness, objectivity, and emotional stability.

Gage (62) demonstrated that logically derived scoring keys for the *MTAI* yielded slightly higher validity and reliability coefficients than the empirically based keys now in use. Although mean scores on the *MTAI* were significantly raised through an educational psychology course, the fact that course content examinations were not significantly correlated with *MTAI* scores caused Eson (50) to question the validity of the *MTAI* for measuring attitudes. Fishman (55) compared *MTAI* scores of various subgroups of teachers based on six factors.

As criteria for a concurrent validity study of the *MTAI*, ratings by pupils and by supervisors were used by Stein and Hardy (120). Raw score *r*'s with the *MTAI* were .43 and .17, but conversion into T scores and combining the criteria yielded an *r* of .56. A test-retest *r* of .92 was obtained for an estimate of reliability.

Edwards Personal Preference Schedule (EPPS)

Although published only five years ago, the *EPPS* has already been used in a large number of investigations, so many that only a sampling of the studies will be reviewed in this section.

Allen (2, 3) obtained *r*'s between the *EPPS* variables and the *MMPI* Scales for 82 college men and 48 women. The intercorrelations among the *MMPI* Scales and the *EPPS* variables were quite low, but substantial *r*'s were found between the *EPPS* variables, suggesting a lack of independence for these alleged independent components. Merrill and Heathers (101) obtained essentially the same results with another college group. Comparisons with *MMPI* and *EPPS* norms were also made for both scales, revealing greater differences from norms for the *MMPI*. The large discrepancy in the case of the *MMPI* might be attributable to the fact that the *MMPI* norm group is the general adult population, not college students.

Gebhart and Hoyt (63) used the *EPPS* to compare over- and underachievers. Significant differences (.05 level) were found for seven of the 16 scales. Overachievers had significantly higher mean scores on *Achieve-*

ment, Order, Introception, and *Consistency,* and underachievers were significantly higher on *Nuturance, Affiliation,* and *Change.* Although these results are interesting and suggestive, cross validation might well show different results.

An important contribution was made by Bernardin and Jessor (17) in their approach to a construct validation of the *EPPS* for measuring dependency. Through the use of three experimental task situations requiring explicit demonstration of independent-dependency behavior, comparisons were made with the inferred *EPPS* dependency scores (*Autonomy* and *Deference* scales). Although the results from the third experimental task were ambiguous, the data from the first two experiments provided support for the construct validity of the *Autonomy* and *Deference* scales of the *EPPS.*

In a cross validation of the social desirability scale values for the *EPPS* with high-school students, Klett (85) obtained an *r* of .94 with those originally derived by Edwards. No differences were obtained among different socioeconomic groups or between grades or sexes.

A dittoed bibliography prepared in August 1958 for the *EPPS* by the Psychological Corporation contained 66 references to the *Schedule,* of which 15 were unpublished theses.

The Strong Vocational Interest Blank (SVIB)

The *Strong Vocational Interest Blank* continued to be a popular instrument for empirical studies. The *Physician* Scale of the *SVIB* did not differentiate between successful and unsuccessful premedical students (99), nor did scores on the *SVIB* correlate significantly with veterinary medicine grades (73). However, the restricted range of high scores on the scale as well as the restricted range of grades may account for the low correlation. A test-retest *r* on the veterinary scale over a four-year period was .68 (73). King (84) used five stability measures, demonstrating that *SVIB* scores remained stable. Scores of Powers' subjects (106) showed permanency over 10 years regardless of age, aptitudes, education, vocational opportunity, or economic status.

McCornack (96) was able to develop separate *SVIB* keys for male and female social workers, and Witkin (140) demonstrated the existence of differential interest patterns in salesmen. Hughes and McNamara (80) developed "custom-built" sales interest keys for accounting and data-processing machine salesmen and electric typewriter salesmen. Dunnette (48) reported a preliminary study on the use of *SVIB* scores to discriminate among engineers engaged in research and development, production, or sales.

Lyerly (94) computed "chance" scores for the different scales of the *SVIB,* and Perry (105) reported that a forced-choice format instead of the L.I.D. format used by the *SVIB* is a superior method for differentiating groups.

Kuder Preference Record (KPR)

Studies with the *KPR* were made to develop profiles for professional forest service men (19), different kinds of psychologists (88), pharmaceutical salesmen (88), and Air Force officers (69). Stewart and Roberts (122) explored the use of the *KPR* to differentiate between students persisting in teacher training and those leaving.

Forer (57) compared *KPR* scores for 36 emotionally and/or physically disabled veterans before and after occupational frustration. He found a significant decrement in social service scores and a significant increase in musical interests.

KPR scores of honors majors in 11 different college departments were compared by Bendig (15), with eight of the nine *KPR* scales significantly discriminating interest differences among the 11 groups. A *Research Handbook* (89) containing a rich reservoir of information on interest measurements was published.

Other Instruments

In addition to those covered above, two instruments which received much attention in the literature were the *California F Scale* and the *Gordon Personal Profile*. Titus and Hollander (134) reviewed the literature from 1950 to 1955 on the *F Scale*, and additional studies were reported (28, 30, 78, 111, 125, 136).

Eight validity studies of the *Gordon Personal Profile* that had not appeared in the general literature are reported in the revised manual (66). Three "susceptibility to faking" studies with the *Gordon Personal Profile* were described previously in this chapter.

Studies involving a number of other established inventories besides those reviewed above have also appeared in the last three years.

Drevdahl and Cattell (47) compared the scores of outstanding artists and writers on the Cattell *16 PF* scale with those of the normative population. Karson and Pool (83) compared scores made by Air Force officers on the Cattell *16 PF* scale with *MMPI* scores and *Wechsler* IQ's. Failure of experienced clinical psychologists and psychiatrists to predict correlations of *16 PF* factors with *MMPI* scores suggested a need to redefine the *PF* factors.

Cuadra and Reed (40) found that the *California Psychological Inventory (CPI)* did not provide a consistent means for predicting psychiatric aide performance. Bennett and Rudoff (16) demonstrated that having items on the *CPI* read aloud did not seriously distort group trends or individual profiles. Bauernfeind (12) investigated the use of an item format permitting expression of strength of response through the *SRA Youth Inventory*.

Barthol and Zeigler (9) used *How Supervise?* to measure gains following a supervisory training program, recognizing that gains in inventory

scores did not provide evidence of validity of the scale. Decker (44) found that scores on *How Supervise?* did not correlate with ratings of supervisors, but item analyses of the scale did show high consistency suggesting internal validity.

Cantoni (24) found that the *Bell Adjustment Inventory* was able to contribute in a multiple correlation for predicting occupational status 10 years after high-school graduation. Singer and Stefflre (114) observed that veterans checking many problems on the *Mooney Problem Check List* tended to have undesirable scores on the *Guilford-Zimmerman Temperament Survey.*

Studies (1, 20) with the *Survey of Study Habits and Attitudes* demonstrated its usefulness in counseling students or in working with them in a reading improvement program. Though a widely used inventory, very little in the way of published research has been available on the *Humm-Wadsworth Temperament Scale.*

A large number of new instruments appeared in the last three years for which not many data are available. Krathwohl and Cronbach (87) described ways in which the *Squares Test* might be used in the measurement of personality. Stone (124) devised a structured multiple-choice version of the *Rorsehach* and presented some preliminary validity data. Schutter and Maher (110) constructed a forced-choice *Study Activity Questionnaire* to predict grades, and Ewens (51) prepared an *Activity Experience Inventory* to measure manifest interest. Symonds (126) developed an *Educational Interest Inventory* with five interest clusters.

Harris (74) constructed *A Scale for Measuring Attitudes of Social Responsibility in Children.* Eysenck (52) reported on the development of a short questionnaire for measuring neuroticism and extroversion. Spector (119) described the *Officer Behavior Description* and an *Attitudes Test in Human Relations.* Siegel (112, 113) described the construction and validation of a *Biographical Inventory for Students,* and Soar (116) used personal history data to predict success in service station management. Bass (11) discussed the validation of a *Proverbs Personality Test.* Bauernfeind (12) dealt with children's strength of response to attitude items.

Summary

This review of the literature has indicated that only a few of the hundreds of published personality inventories have been used to accumulate an appreciable amount of data. In general, each inventory has had very little attention in the literature except through the efforts of its author. It is to be lamented that so much research effort has been dissipated on such a multitude of instruments instead of being focused on a promising few. It also is evident that none of the inventories has adequate normative data, the *MMPI* probably coming closest to having adequate norms.

Values and attitudes in different subgroups will make norms derived from other groups of different backgrounds inappropriate; thus, the user of personality inventories must be prepared to develop his own normative data for the group with which he is working.

Bibliography

1. AHMANN, J. S., and GLOCK, MARVIN D. "The Utility of a Study Habits and Attitudes Inventory in a College Reading Program." *Journal of Educational Research* 51: 297-303; December 1957.
2. ALLEN, ROBERT M. "An Analysis of Edwards Personal Preference Schedule, Intercorrelations for a Local College Population." *Journal of Educational Research* 51: 591-97; April 1958.
3. ALLEN, ROBERT M. "The Relationship Between the Edwards Personal Preference Schedule Variables and the Minnesota Multiphasic Personality Inventory Scales." *Journal of Applied Psychology* 41: 307-11; October 1957.
4. AUBLE, DONAVON. "Validity Indices for the Heston Personal Adjustment Inventory." *Journal of Applied Psychology* 41: 79-81; April 1957.
5. BARNES, EUGENE H. "Factors, Response Bias, and the MMPI." *Journal of Consulting Psychology* 20: 419-21; December 1956.
6. BARNES, EUGENE H. "Response Bias and the MMPI." *Journal of Consulting Psychology* 20: 371-74; October 1956.
7. BARR, ARVIL S., and JONES, ROBERT E. "The Measurement and Prediction of Teaching Efficiency." *Review of Educational Research* 28: 256-64; June 1958.
8. BARTHOL, RICHARD P., and KIRK, BARBARA A. "The Selection of Graduate Students in Public Health Education." *Journal of Applied Psychology* 40: 159-63; June 1956.
9. BARTHOL, RICHARD P., and ZEIGLER, MARTIN. "Evaluation of a Supervisory Training Program with *How Supervise?*" *Journal of Applied Psychology* 40: 403-405; December 1956.
10. BASS, BERNARD M. "Faking by Sales Applicants of a Forced Choice Personality Inventory." *Journal of Applied Psychology* 41: 403-404; December 1957.
11. BASS, BERNARD M. "Validity Studies of a Proverbs Personality Test." *Journal of Applied Psychology* 41: 158-60; June 1957.
12. BAUERNFEIND, ROBERT H. "Measuring Children's Strength of Response to Attitude Items." *Educational and Psychological Measurement* 15: 63-70; Spring 1955.
13. BELL, GRAHAM B., and STOLPER, RHODA. "An Attempt at Validation of the Empathy Test." *Journal of Applied Psychology* 39: 442-43; December 1955.
14. BENDIG, ALBERT W. "The Development of a Short Form of the Manifest Anxiety Scale." *Journal of Consulting Psychology* 20: 384; October 1956.
15. BENDIG, ALBERT W. "Validity of Kuder Differences Among Honors Majors." *Educational and Psychological Measurement* 17: 593-98; Winter 1957.
16. BENNETT, LAWRENCE A., and RUDOFF, ALVIN. "Evaluation of Modified Administration of the California Psychological Inventory." *Journal of Clinical Psychology* 13: 303-304; July 1957.
17. BERNARDIN, ALFRED C., and JESSOR, RICHARD. "A Construct Validation of the Edwards Personal Preference Schedule with Respect to Dependency." *Journal of Consulting Psychology* 21: 63-67; February 1957.
18. BORISLOW, BERNARD. "The Edwards Personal Preference Schedule (EPPS) and Fakability." *Journal of Applied Psychology* 42: 22-27; February 1958.
19. BRODY, DAVID S. "Kuder Interest Patterns of Professional Forest Service Men." *Educational and Psychological Measurement* 17: 599-605; Winter 1957.
20. BROWN, WILLIAM F., and HOLTZMAN, WAYNE H. "Use of the Survey of Study Habits and Attitudes for Counseling Students." *Personnel and Guidance Journal* 35: 215-18; December 1956.
21. BUSS, ARNOLD H. "A Follow-Up Item Analysis of the Taylor Anxiety Scale." *Journal of Clinical Psychology* 11: 409-10; October 1955.
22. CALVIN, ALLEN D., and OTHERS. "A Further Investigation of the Relationship Between Manifest Anxiety and Intelligence." *Journal of Consulting Psychology* 19: 280-82; August 1955.

23. CALVIN, ALLEN D., and HANLEY, CHARLES. "An Investigation of Dissimulation on the MMPI by Means of the 'Lie Detector.'" *Journal of Applied Psychology* 41: 312-16; October 1957.
24. CANTONI, LOUIS J. "High School Tests and Measurements as Predictors of Occupational Status." *Journal of Applied Psychology* 39: 253-55; August 1955.
25. CASTANEDA, ALFRED; MCCANDLESS, BOYD R.; and PALERMO, DAVID S. "The Children's Form of the Manifest Anxiety Scale." *Child Development* 27: 317-27; September 1956.
26. CASTANEDA, ALFRED; PALERMO, DAVID S.; and MCCANDLESS, BOYD R. "Complex Learning and Performance as a Function of Anxiety in Children and Task Difficulty." *Child Development* 27: 327-33; September 1956.
27. CATTELL, RAYMOND B. "Validation and Intensification of the Sixteen Personality Factor Questionnaire." *Journal of Clinical Psychology* 12: 205-14; July 1956.
28. CHAPMAN, LOREN J., and CAMPBELL, DONALD T. "Response Set in the F Scale." *Journal of Abnormal and Social Psychology* 54: 129-32; January 1957.
29. CHRISTIE, RICHARD, and BUDNITZKY, STANLEY. "A Short Forced Choice Anxiety Scale." *Journal of Consulting Psychology* 21: 501; December 1957.
30. COHN, THOMAS S. "The Relation of the F Scale to a Response Set To Answer Positively." *Journal of Social Psychology* 44: 129-33; August 1956.
31. COMREY, ANDREW L. "A Factor Analysis of Items on the MMPI Depression Scale." *Educational and Psychological Measurement* 17: 578-85; Winter 1957.
32. COMREY, ANDREW L. "A Factor Analysis of Items on the MMPI Hypochondriasis Scale." *Educational and Psychological Measurement* 17: 568-77; Winter 1957.
33. COMREY, ANDREW L. "A Factor Analysis of Items on the MMPI Hypomania Scale." *Educational and Psychological Measurement* 18: 313-23; Summer 1958.
34. COMREY, ANDREW L. "A Factor Analysis of Items on the MMPI Hysteria Scale." *Educational and Psychological Measurement* 17: 586-92; Winter 1957.
35. COMREY, ANDREW L. "A Factor Analysis of Items on the MMPI Paranoia Scale." *Educational and Psychological Measurement* 18: 99-107; Spring 1958.
36. COMREY, ANDREW L. "A Factor Analysis of Items on the MMPI Psychasthenia Scale." *Educational and Psychological Measurement* 18: 293-300; Summer 1958.
37. COMREY, ANDREW L. "A Factor Analysis of Items on the MMPI Psychopathic Deviate Scale." *Educational and Psychological Measurement* 18: 91-98; Spring 1958.
38. COMREY, ANDREW L., and MARGGRAFF, WALTRAUD M. "A Factor Analysis of Items on the MMPI Schizophrenia Scale." *Educational and Psychological Measurement* 18: 301-11; Summer 1958.
39. CRONBACH, LEE J., and GLESER, GOLDINE C. *Psychological Tests and Personnel Decisions.* Urbana: University of Illinois Press, 1957. 165 p.
40. CUADRA, CARLOS A., and REED, CHARLES F. "Prediction of Psychiatric Aide Performance." *Journal of Applied Psychology* 41: 195-97; June 1957.
41. DAVID, HENRY P., and VON BRACKEN, HELMUT, editors. *Perspectives in Personality Theory.* New York: Basic Books, 1957. 435 p.
42. DAVIDS, ANTHONY. "Relations Among Several Objective Measures of Anxiety Under Different Conditions of Motivation." *Journal of Consulting Psychology* 19: 275-79; August 1955.
43. DAVIDS, ANTHONY, and ERIKSEN, CHARLES W. "The Relation of Manifest Anxiety to Association Productivity and Intellectual Attainment." *Journal of Consulting Psychology* 19: 219-22; June 1955.
44. DECKER, ROBERT L. "An Item Analysis of *How Supervise?* Using Both Internal and External Criteria." *Journal of Applied Psychology* 40: 406-11; December 1956.
45. DELLA PIANA, GABRIEL M., and GAGE, NATHANIEL L. "Pupils' Values and the Validity of the Minnesota Teacher Attitude Inventory." *Journal of Educational Psychology* 46: 167-78; March 1955.
46. DREGER, RALPH M., and AIKEN, LEWIS R., JR. "The Identification of Number Anxiety in a College Population." *Journal of Educational Psychology* 48: 344-51; October 1957.
47. DREVDAHL, JOHN E., and CATTELL, RAYMOND B. "Personality and Creativity in Artists and Writers." *Journal of Clinical Psychology* 14: 107-11; April 1958.

48. DUNNETTE, MARVIN D. "Vocational Interest Differences Among Engineers Employed in Different Functions." *Journal of Applied Psychology* 41: 273-78; October 1957.
49. ELLIS, ALBERT. "The Validity of Personality Questionnaires." *Psychological Bulletin* 43: 385-440; September 1946.
50. ESON, MORRIS E. "The Minnesota Teacher Attitude Inventory in Evaluating the Teaching of Educational Psychology." *Journal of Educational Psychology* 47: 271-75; May 1956.
51. EWENS, WILLIAM P. "The Development and Standardization of a Preliminary Form of an Activity Experience Inventory: A Measure of Manifest Interest." *Journal of Applied Psychology* 40: 169-74; June 1956.
52. EYSENCK, HANS J. "A Short Questionnaire for the Measurement of Two Dimensions of Personality." *Journal of Applied Psychology* 42: 14-17; February 1958.
53. FARBER, ISADORE E., and SPENCE, KENNETH W. "Effects of Anxiety, Stress, and Task Variables on Reaction Time." *Journal of Personality* 25: 1-18; September 1956.
54. FIEDLER, FRED E., and OTHERS. "Interrelations Among Measures of Personality Adjustment in Nonclinical Populations." *Journal of Abnormal and Social Psychology* 56: 345-51; May 1958.
55. FISHMAN, JOSHUA A. "The MTAI in an American Minority-Group School Setting: I. Differences Between Test Characteristics for Norm and Non-norm Populations." *Journal of Educational Psychology* 48: 41-51; January 1957.
56. FORDYCE, WILBERT E. "Social Desirability in the MMPI." *Journal of Consulting Psychology* 20: 171-75; June 1956.
57. FORER, BERTRAM R. "The Stability of Kuder Scores in a Disabled Population." *Educational and Psychological Measurement* 15: 166-69; Summer 1955.
58. FRENCH, ELIZABETH G. "A Note on the Edwards Personal Preference Schedule for Use with Basic Airmen." *Educational and Psychological Measurement* 18: 109-15; Spring 1958.
59. FRICK, JAMES W., and KEENER, HELEN E. "A Validation Study of the Prediction of College Achievement." *Journal of Applied Psychology* 40: 251-52; August 1956.
60. FRICKE, BENNO G. "Response Set as a Suppressor Variable in the OAIS and MMPI." *Journal of Consulting Psychology* 20: 161-69; June 1956.
61. FURST, EDWARD J., and FRICKE, BENNO G. "Development and Applications of Structured Tests of Personality." *Review of Educational Research* 26: 26-55; February 1956.
62. GAGE, NATHANIEL L. "Logical Versus Empirical Scoring Keys: The Case of the MTAI." *Journal of Educational Psychology* 48: 213-16; April 1957.
63. GEBHART, G. GARY, and HOYT, DONALD P. "Personality Needs of Under- and Overachieving Freshmen." *Journal of Applied Psychology* 42: 125-28; April 1958.
64. GEHMAN, WINFIELD S. "A Study of Ability To Fake Scores on the Strong Vocational Interest Blank for Men." *Educational and Psychological Measurement* 17: 65-70; Spring 1957.
65. GOODLING, RICHARD A. "Relationship Between the IM Scale of the SVIB and Scales of the Guilford-Zimmerman." *Journal of Counseling Psychology* 3: 146-49; Summer 1956.
66. GORDON, LEONARD V. *Gordon Personal Profile.* Yonkers-on-Hudson, N. Y.: World Book Co., 1958.
67. GORDON, LEONARD V., and STAPLETON, ERNEST S. "Fakability of a Forced-Choice Personality Test Under Realistic High School Employment Conditions." *Journal of Applied Psychology* 40: 258-62; August 1956.
68. GOWAN, JOHN C., and GOWAN, MAY S. "A Teacher Prognosis Scale for the MMPI." *Journal of Educational Research* 49: 1-12; September 1955.
69. GUBA, EGON E., and GETZELS, JACOB W. "Interest and Value Patterns of Air Force Officers." *Educational and Psychological Measurement* 16: 465-70; Winter 1956.
70. GUILFORD, JOY P., and ZIMMERMAN, WAYNE S. *Fourteen Dimensions of Temperament.* Psychological Monographs, No. 417. Washington, D. C.: American Psychological Association, 1956. 26 p.

71. HALL, CALVIN S., and LINDZEY, GARDNER. *Theories of Personality*. New York: John Wiley and Sons, 1957. 572 p.
72. HANLEY, CHARLES. "Social Desirability and Responses to Items from Three MMPI Scales: *D*, *Sc*, and *K*." *Journal of Applied Psychology* 40: 324-28; October 1956.
73. HANNUM, THOMAS E., and THRALL, JOHN B. "Use of the Strong Vocational Interest Blank for Prediction in Veterinary Medicine." *Journal of Applied Psychology* 39: 249-52; August 1955.
74. HARRIS, DALE B. "A Scale for Measuring Attitudes of Social Responsibility in Children." *Journal of Abnormal and Social Psychology* 55: 322-26; November 1957.
75. HERON, ALASTAIR. "The Effects of Real-Life Motivation on Questionnaire Response." *Journal of Applied Psychology* 40: 65-68; April 1956.
76. HERON, ALASTAIR. "The Objective Assessment of Personality Among Female Unskilled Workers." *Educational and Psychological Measurement* 15: 117-26; Summer 1955.
77. HEWER, VIVIAN H. "A Comparison of Successful and Unsuccessful Students in the Medical School at the University of Minnesota." *Journal of Applied Psychology* 40: 164-68; June 1956.
78. HINES, VYNCE A. "F Scale, GAMIN, and Public School Principal Behavior." *Journal of Educational Psychology* 47: 321-28; October 1956.
79. HOLZ, WILLIAM C.; HARDING, GEORGE F.; and GLASSMAN, SIDNEY M. "A Note on the Clinical Validity of the Marsh-Hilliard-Liechti MMPI Sexual Deviation Scale." *Journal of Consulting Psychology* 21: 326; August 1957.
80. HUGHES, JOHN L., and MCNAMARA, WALTER J. "Limitations on the Use of Strong Sales Keys for Selection and Counseling." *Journal of Applied Psychology* 42: 93-96; April 1958.
81. IZARD, CARROLL E., and ROSENBERG, NATHAN. "Effectiveness of a Forced-Choice Leadership Test Under Varied Experimental Conditions." *Educational and Psychological Measurement* 18: 57-62; Spring 1958.
82. KAESS, WALTER A., and WITRYOL, SAM L. "Positive and Negative Faking on a Forced-Choice Authoritarian Scale." *Journal of Applied Psychology* 41: 333-39; October 1957.
83. KARSON, SAMUEL, and POOL, KENNETH B. "The Construct Validity of the Sixteen Personality Factors Test." *Journal of Clinical Psychology* 13: 245-52; July 1957.
84. KING, LESLIE A. "Stability Measures of Strong Vocational Interest Blank Profiles." *Journal of Applied Psychology* 41: 143-47; June 1957.
85. KLETT, C. JAMES. "The Stability of the Social Desirability Scale Values in the Edwards Personal Preference Schedule." *Journal of Consulting Psychology* 21: 183-85; April 1957.
86. KLUGMAN, SAMUEL F. "A Study of the Interest Profile of a Psychotic Group and Its Bearing on Interest-Personality Theory." *Educational and Psychological Measurement* 17: 55-64; Spring 1957.
87. KRATHWOHL, DAVID R., and CRONBACH, LEE. "Suggestions Regarding a Possible Measure of Personality: The Squares Test." *Educational and Psychological Measurement* 16: 305-16; Autumn 1956.
88. KUDER, G. FREDERIC. "A Comparative Study of Some Methods of Developing Occupational Keys." *Educational and Psychological Measurement* 17: 105-14; Spring 1957.
89. KUDER, G. FREDERIC. *Kuder Preference Record Occupational Research Handbook*. Chicago: Science Research Associates, 1957.
90. LABUE, ANTHONY C. "Personality Traits and Persistence of Interest in Teaching as a Vocational Choice." *Journal of Applied Psychology* 39: 362-65; October 1955.
91. LEEDS, CARROLL H. "Teacher Attitudes and Temperament as a Measure of Teacher-Pupil Rapport." *Journal of Applied Psychology* 40: 333-37; October 1956.
92. LEWIS, NAN A., and TAYLOR, JANET A. "Anxiety and Extreme Response Preferences." *Educational and Psychological Measurement* 15: 111-16; Summer 1955.
93. LOEVINGER, JANE. "Some Principles of Personality Measurement." *Educational and Psychological Measurement* 15: 3-17; Spring 1955.
94. LYERLY, SAMUEL B. "'Chance' Scores on the Strong Vocational Interest Blank for Men." *Journal of Applied Psychology* 41: 141-42; June 1957.

95. MCCANDLESS, BOYD R.; CASTANEDA, ALFRED; and PALERMO, DAVID S. "Anxiety in Children and Social Status." *Child Development* 27: 385-92; December 1956.
96. MCCORNACK, ROBERT L. "Vocational Interests of Male and Female Social Workers." *Journal of Applied Psychology* 40: 11-13; February 1956.
97. MARTIN, BARCLAY, and MCGOWAN, BRUCE. "Some Evidence on the Validity of the Sarason Test Anxiety Scale." *Journal of Consulting Psychology* 19: 468; December 1955.
98. MATARAZZO, JOSEPH D. "MMPI Validity Scores as a Function of Increasing Levels of Anxiety." *Journal of Consulting Psychology* 19: 213-17; June 1955.
99. MELTON, RICHARD S. "Differentiation of Successful and Unsuccessful Premedical Students." *Journal of Applied Psychology* 39: 397-400; December 1955.
100. MELTON, WILLIAM R., JR. "An Investigation of the Relationship Between Personality and Vocational Interest." *Journal of Educational Psychology* 47: 163-74; March 1956.
101. MERRILL, REED M., and HEATHERS, LOUISE B. "The Relation of the MMPI to the Edwards Personal Preference Schedule on a College Counseling Center Sample." *Journal of Consulting Psychology* 20: 310-14; August 1956.
102. MITZEL, HAROLD E., and OTHERS. "The Effects of Response Sets on the Validity of the Minnesota Teacher Attitude Inventory." *Educational and Psychological Measurement* 16: 501-15; Winter 1956.
103. PATTERSON, CECIL H. "Interest Tests and the Emotionally Disturbed Client." *Educational and Psychological Measurement* 17: 264-80; Summer 1957.
104. PEEK, ROLAND M., and STORMS, LOWELL H. "Validity of the Marsh-Hilliard-Liechti MMPI Sexual Deviation Scale in a State Hospital Population." *Journal of Consulting Psychology* 20: 133-36; April 1956.
105. PERRY, DALLIS K. "Forced-Choice vs. L-I-D Response Items in Vocational Interest Measurement." *Journal of Applied Psychology* 39: 256-62; August 1955.
106. POWERS, MABEL K. "Permanence of Measured Vocational Interests of Adult Males." *Journal of Applied Psychology* 40: 69-72; April 1956.
107. ROSEN, HJALMAR, and ROSEN, R. A. HUDSON. "Personality Variables and Role in a Union Business Agent Group." *Journal of Applied Psychology* 41: 131-36; April 1957.
108. RUSMORE, JAY T. "Fakability of the Gordon Personal Profile." *Journal of Applied Psychology* 40: 175-77; June 1956.
109. SCHULZ, R. E., and CALVIN, ALLEN D. "A Failure To Replicate the Finding of a Negative Correlation Between Manifest Anxiety and ACE Scores." *Journal of Consulting Psychology* 19: 223-24; June 1955.
110. SCHUTTER, GENEVIEVE, and MAHER, HOWARD. "Predicting Grade-Point Average with a Forced-Choice Study Activity Questionnaire." *Journal of Applied Psychology* 40: 253-57; August 1956.
111. SHELLEY, HARRY P. "Response Set and the California Attitude Scales." *Educational and Psychological Measurement* 16: 63-67; Spring 1956.
112. SIEGEL, LAURENCE. "A Biographical Inventory for Students: 1. Construction and Standardization of the Instrument." *Journal of Applied Psychology* 40: 5-10; February 1956.
113. SIEGEL, LAURENCE. "A Biographical Inventory for Students: II. Validation of the Instrument." *Journal of Applied Psychology* 40: 122-26; April 1956.
114. SINGER, STANLEY L., and STEFFLRE, BUFORD. "The Concurrent Validity of the Mooney Problem Check List." *Personnel and Guidance Journal* 35: 298-301; January 1957.
115. SINICK, DANIEL. "Two Anxiety Scales Correlated and Examined for Sex Differences." *Journal of Clinical Psychology* 12: 394-95; October 1956.
116. SOAR, ROBERT S. "Personal History Data as a Predictor of Success in Service Station Management." *Journal of Applied Psychology* 40: 383-85; December 1956.
117. SORENSON, A. GARTH. "A Note on the 'Fakability' of the Minnesota Teacher Attitude Inventory." *Journal of Applied Psychology* 40: 192-94; June 1956.
118. SORENSON, A. GARTH, and SHELDON, MARTIN S. "A Further Note on the Fakability of the MTAI." *Journal of Applied Psychology* 42: 74-78; April 1958.
119. SPECTOR, AARON J. "Human Relations Behavior on the Job: The Officer Behavior Description." *Journal of Applied Psychology* 41: 110-13; April 1957.

120. STEIN, HARRY L., and HARDY, JAMES. "A Validation Study of the Minnesota Teacher Attitude Inventory in Manitoba." *Journal of Educational Research* 50: 321-38; January 1957.
121. STERN, GEORGE G.; STEIN, MORRIS I.; and BLOOM, BENJAMIN S. *Methods in Personality Assessment.* Glencoe, Ill.: Free Press, 1956. 271 p.
122. STEWART, LAWRENCE H., and ROBERTS, JOSEPH P. "The Relationship of Kuder Profiles to Remaining in a Teachers' College and to Occupational Choice." *Educational and Psychological Measurement* 15: 416-21; Winter 1955.
123. STEWART, ROGER G. "Factor Analysis in the Measurement of Personality Integration." *Educational and Psychological Measurement* 16: 471-80; Winter 1956.
124. STONE, JOICE B. *S-O Rorschach Test Manual.* Los Angeles: California Test Bureau, 1958.
125. SUNDBERG, NORMAN D., and BACHELIS, WARREN D. "The Fakability of Two Measures of Prejudice: The California *F* Scale and Gough's *Pr* Scale." *Journal of Abnormal and Social Psychology* 52: 140-42; January 1956.
126. SYMONDS, PERCIVAL M. "An Educational Interest Inventory." *Educational and Psychological Measurement* 18: 377-85; Summer 1958.
127. TAFT, RONALD. "A Cross-Cultural Comparison of the MMPI." *Journal of Consulting Psychology* 21: 161-64; April 1957.
128. TAMKIN, ARTHUR S. "An Evaluation of the Construct Validity of Barron's Ego-Strength Scale." *Journal of Clinical Psychology* 13: 156-58; April 1957.
129. TAULBEE, EARL S. "A Validation of MMPI Scale Pairs in Psychiatric Diagnosis." *Journal of Clinical Psychology* 14: 316; July 1958.
130. TAYLOR, JANET A. "Drive Theory and Manifest Anxiety." *Psychological Bulletin* 53: 303-20; July 1956.
131. TAYLOR, JANET A. "The Effects of Anxiety Level and Psychological Stress on Verbal Learning." *Journal of Abnormal and Social Psychology* 57: 55-60; July 1958.
132. TESSENEER, RALPH, and TYDLASKA, MARY. "A Cross-Validation of a Work Attitude Scale from the MMPI." *Journal of Educational Psychology* 47: 1-7; January 1956.
133. TINDALL, RALPH H. "Relationships Among Indices of Adjustment Status." *Educational and Psychological Measurement* 15: 152-62; Summer 1955.
134. TITUS, H. EDWIN, and HOLLANDER, EDWIN P. "The California F Scale in Psychological Research: 1950-1955." *Psychological Bulletin* 54: 47-64; January 1957.
135. VOAS, ROBERT B. "A Procedure for Reducing the Effects of Slanting Questionnaire Responses Toward Social Acceptability." *Educational and Psychological Measurement* 18: 337-45; Summer 1958.
136. WELLS, WILLIAM D.; CHIARAVALLO, G.; and GOLDMAN, S. "Brothers Under the Skin: A Validity Test of the F Scale." *Journal of Social Psychology* 45: 35-40; February 1957.
137. WELSH, GEORGE S., and DAHLSTROM, W. GRANT, editors. *Basic Readings on the MMPI in Psychology and Medicine.* Minneapolis: University of Minnesota Press, 1956. 656 p.
138. WINTER, WILLIAM D., and SALCINES, RAMON A. "The Validity of the Objective Rorschach and the MMPI." *Journal of Consulting Psychology* 22:199-202; June 1958.
139. WIRT, ROBERT D., and BROEN, WILLIAM E., JR. "The Relation of the Children's Manifest Anxiety Scale to the Concept of Anxiety as Used in the Clinic." *Journal of Consulting Psychology* 20: 482; December 1956.
140. WITKIN, ARTHUR A. "Differential Interest Patterns in Salesmen." *Journal of Applied Psychology* 40: 338-40; October 1956.

CHAPTER VI

Development and Applications of Projective Techniques

ROBERT A. HEIMANN and JOHN W. M. ROTHNEY

THE AUTHORS wrote in a REVIEW article three years ago (38) that research with projective techniques presented more of a challenge than research with conventional psychometric methods because there was no clear-cut agreement as to the rationale for the whole process, because it was extremely difficult to find reliable criterion measures, and because there was no common metric. As was indicated, due to these difficulties the researcher with projective methods often failed to employ scientific methods, failed to use control groups, used too few cases, tended to overgeneralize his findings, described his scoring procedures too vaguely, used ill-defined criterion measures, and continued to rework concepts that research had shown to be neither important nor meaningful. It seems to the writers that most of the research reports in the period covered by this review are still limited by such difficulties.

One of the leading Rorschachers, Klopfer, described the current situation when he said that there seemed to be a dearth of carefully designed longitudinal studies, and the evaluation of published studies seemed to be a thankless and almost impossible task because of the multiplicity of scoring systems which are not mutually translatable (22). He also charged that findings are so influenced by the researcher's choice of method of administration and scoring that any comparison of results seemed almost impossible. Two major critical reports are worth special consideration. Cronbach and Meehl (8) stressed the need to use widespread negative evidence so often found in studies with projective techniques and suggested that these results might contribute in building psychological constructs which would add to the validity of projective techniques as a whole. Lindzey (29) pointed out that one of the factors which contributed to the slight progress made toward an understanding of the *Thematic Apperception Test (TAT)* was an excess of casual empiricism and a scarcity of systematic investigations.

Validity and Reliability Studies

The concepts of construct and concurrent validity are widely used in reports which attempt to establish the validity of projective techniques. Few studies used predictive validity designs. In the sample of research which follows, the authors have not attempted to cover the entire field or to report on the numerous published studies with great completeness.

Those selected seemed to illustrate particular methodologies. They were deemed representative of current research in this field.

In the period under review the bulk of validation and reliability studies was reported for the *Rorschach*. Wysocki (52) described the preferences of his 374 adult subjects for particular *Rorschach* cards in a carefully designed study. He ranked the 10 *Rorschach* cards on the basis of these ranks. Further analysis was made of the popular choices of 13 groups chosen on the basis of sex, IQ, "adjustment," and other variables. Low and moderate coefficients were offered as evidence that popularity of card choices was a useful diagnostic aid. Sommer and Sommer (46) tried to identify aggressive behavior by rating color responses of male veteran patients to the *Rorschach* and by designating these color responses as aggressive or not. When the color aggressive responses were compared with behavioral evidence of aggression taken from case histories of the 57 subjects, a better-than-chance relationship was found. The authors concluded, however, that the relationships found were not of sufficient magnitude to permit their use in individual prediction of assaultative behavior. Levine, Glass, and Meltzoff (26) divided a group of 274 outpatient veterans into two groups on the basis of whether or not they reversed the letter "N" on the digit symbol subtest of the *Wechsler-Bellevue Scales*. They hypothesized that reversals on this item suggested inhibition in ego function and should be accompanied by more movement responses on the *Rorschach*. The findings supported their hypothesis beyond the chance level.

Traditionally the most common method of seeking validation evidence with personality measurement has been that of comparing the test performance of two or more groups of known characteristics. A question might be raised about whether significant differences found in preclassified groups is a demonstration of the validity of a measure unless it is checked by adequate cross validation. Griffin (17) in a concurrent validity study examined the hypothesis that there is a positive relationship between creativity and movement responses on the *Rorschach*. She administered this instrument to two groups of 20 college women who were classified as creative or not by their teachers. These groups were matched by age, year in college, and scores on the *ACE Psychological Examination for College Freshmen*. Care was taken to minimize scoring errors by having eight scorers, but the sample was very small and definitions of creativity lacked clarity. Her findings did not support the classic *Rorschach* supposition that more M responses are associated with greater creativity.

Despite consistent negative findings, studies utilizing "blind" readings of *Rorschachs* still appear. Chambers and Hamlin (6) found 20 experienced clinicians correct 58 times out of 100 in their attempts to classify five *Rorschach* protocols into proper diagnostic categories. The clinicians were successful in identifying mental defectives 90 percent of the time although this might seem to be a rather empty claim since there

are several established intelligence tests available for this kind of classification. Shaw and Cruickshank (43) used the *Rorschach* with two groups of 25 children in an attempt to prove this instrument's efficiency as a diagnostic tool. One of their groups was composed of normal but institutionalized children; the other was composed of children with *grand mal.* They found that the *Rorschach* did not appear to be a useful clinical aid in the diagnosis of idiopathic epilepsy. The *Rorschach* was not useful in differentiating children with defective hearing from those with normal hearing according to Fiedler and Stone (13) who studied 10 matched pairs of such children. Levitt (27) analyzed the *Rorschachs* of 39 disturbed school children and 155 normal controls. Of nine significant differences found in the protocols of these two groups, only one, the shading response, was in the direction predicted. *Rorschach* summary scores cannot be regarded as effective in differentiating psychiatric groups, according to Knopf (23) who studied the *Rorschach* records of 131 psychoneurotics, 106 psychopaths, and 100 patients classified as schizophrenic.

Few studies appeared during the period under review in which the criterion measure was made at some subsequent time in an effort to determine the predictive validity of projective techniques. Cartwright (4) studied the *Rorschach Prognostic Rating Scale (PRS)* scoring technique with 13 cases who were rated before and after nondirective therapy. She concluded that the pretherapy *PRS* score predicted success of therapy, as measured by the therapist's rating of the case, at a better-than-chance level. The very small sample and the possibility of contamination of the criterion rating by the ego-involved counselor make the findings questionable. A carefully designed study was attempted by Eschenbach and Borgatta (12). They compared the objective *Rorschach* scores of 125 airmen classified as normals with their behavior in four sessions in role playing and free discussion situations. Careful ratings were made of the objective behavior of the subjects, and *Rorschach* variables were found not related to these behavioral criteria above a chance level.

It has often been stated that while the *Rorschach* and *TAT* operate at slightly different levels of consciousness, some concurrence of results might be expected. Shatin (42) explored the relationships between these two techniques in a study of 90 hospitalized veteran subjects. Forty *TAT* and 39 *Rorschach* scoring categories were obtained and the chi-square was used to test the significance of the agreement. Of the 1560 null hypotheses tested, 73 were rejected at the .01 level, and 264 at the .05 level of confidence. The author maintained that there was considerable relationship over a wide range of variables from both instruments. The reviewers could draw somewhat different conclusions from the fact that there was approximately 20-percent agreement between the two tests. It would be difficult to predict or to diagnose by one method on the basis of the results of the other. The clinicians' subjective judgments seem to be the main factor in deciding which instrument to use. Davids and others (10) studied

the *TAT* protocols of 20 male paid college-student subjects and concluded that although the *TAT* as a measure of inward-directed aggression did not appear to be valid, it did make a unique contribution to case records. They stated that the unique advantage of projective techniques becomes evident when one examines a given subject's need for expression of intra-aggression. Lindzey and Tejessy (30) analyzed the *TAT* protocols of 20 college students and found aggression scores related to indexes of aggression which were based upon multiple ratings obtained by observers, self, and group of raters. Self-ratings of aggression were most closely associated with *TAT* indications of aggression; the coefficients ranged from .02 to .73. Jensen (19) investigated the relationship of aggressive *TAT* themes to overt behavior in a carefully designed study and found very little concurrence. Of the three groups of high-school boys he used as subjects, he found that those who habitually acted-out aggressively in ways regarded as taboo in schools responded also to the *TAT* with socially taboo content and language.

Two studies used the *TAT* in new ways. Jones (20) reported on the negation *TAT*, an effort to stimulate subjects to produce the most unlikely stories they could. The criterion measures were the reports of therapists who said they felt that the negative stories were more suggestive of repressed psychic content. Lebo and Harrigan (25) instructed 32 female college students to respond to *TAT* pictures in the usual manner and then read them the description of the pictures from the manual. They found that essentially similar protocols were elicited in this visual and aural manner. Product-moment coefficients of the order of .79 between story mood and .76 for level of response were reported for the two procedures.

Several validation studies were concerned with some of the lesser used projective techniques. Churchill and Crandall (7) studied the validity of the *Rotter Incomplete Sentences Test*. They found interscorer reliabilities above .90 and test-retest reliability coefficients in the range of .70 when they tested college-student groups over a period of six months to three years. They also attempted to determine whether this technique would be useful in diagnosis and reported biserial coefficients of .50 with entrance to counseling as their criterion of maladjustment. Boyd and Mandler (3) investigated the concept that children are more apt to respond at length to animal pictures than human pictures; their population was composed of 96 third-grade public-school children. The results of this study suggested that children respond more to human picture stimuli but that the animal cards elicited more original material. Sipprelle and Swensen (45) failed to find evidence to substantiate a favorite hypothesis of users of the *Draw-a-Person Tests* about the meaning of the sex of the figures drawn by subjects. With 49 psychotherapy cases they found no significant relationship between the client's sexual adjustment, as evaluated by his therapist, and sexual characteristics of his human drawings. Silverstein and

Robinson (44) found 75 percent of orthopedically disabled children representing their disability in their drawings; but when they were equated with a comparable normal group, the authors found that they were unable to differentiate the drawings of the disabled children above a chance level.

Much of the difficulty in arriving at satisfactory levels of confidence in validity studies with projective techniques is due to the large error of measurement involved. It seems difficult to assess the influence of the rater upon the rating in the scoring of these instruments, and the reliability of the instrument itself is difficult to ascertain. Datel and Gengerelli (9) attempted to assess the reliability of scorer judgments and interpretations with the *Rorschach*. They found that when 27 well-known clinical psychologists were asked to score and interpret six "blind" *Rorschach* protocols, they were less than successful in matching sets of their reports with those of other psychologists. Eighteen of the 27 judges achieved no more than chance matchings. The authors of this study concluded that a substantial majority of *Rorschach* reports have very little communication value, but that there is a minority which do have significant and adequate interjudge reliability. Lisansky (31) submitted 40 *Rorschach* protocols to six experienced examiners and asked each to answer 10 questions about the subjects' personality, level of intelligence, and adequacy of adjustment. For comparison, life histories of the same subjects were presented to six other clinicians who were asked the same 10 questions. Results suggested that the Rorschachers did not show significantly better agreement than did the judges using life histories only. Hafner (18) explored the influence of time upon *Rorschach* responses of 60 college psychology students. A control group was given the *Rorschach* in the usual manner; the experimental group was instructed to answer as quickly as possible and was limited to two responses per card. Results indicated significant differences between the scoring responses of the two groups and suggested that consideration be given to the time factor in *Rorschach* interpretation.

Attempts to assess the reliability of the *Rorschach* using test-retest methods were tried by Epstein, Nelson, and Tanofsky (11) with a population of 16 college students. They repeated the administration of 10 *Rorschach* cards 10 times over a period of five weeks and found reliability estimates ranging from .20 to .56 for various response categories. They concluded that while all scores measured individual differences to a significant degree, the obtained reliability coefficients were too low for individual use. Rohrer and others (37) determined test-retest reliabilities of group *Rorschach* scoring procedures with 1374 servicemen as subjects. They found one-third of the reliability estimates at .85 or higher and one-third below .56. Seven of the major scoring categories had reliability coefficients of .90 or more. They concluded that their particular group method with objective scoring plus individual inquiry was superior to individual administration and scoring procedures.

Fine (14) attempted to provide a new objective scoring system for the *TAT* and stated that reliability estimates as high as .80 to .91 were obtained when protocols were rated by six graduate students who used his new scoring method. Claims for a satisfactory level of concurrence were made by More (35) who tested 63 pharmacists selected for similar backgrounds in experience, age, and education; More used interviews, biographical summaries, a sentence-completion test, and a shortened form of the *TAT*. He indicated that the congruence among these methods was not so high as that found in reliability studies with objective tests, but stated that when judgments from these different instruments were combined, there was enough congruence for practical use.

Several investigators turned their attention to the problem of establishing the reliability of lesser known projective techniques. A thorough discussion of the problems of establishing satisfactory levels of reliability for projective instruments in view of changes in personality was included by Granick and Scheflen (16) in their study of the *Blacky Pictures*. They administered this instrument to 40 school children and found moderate coefficients of reliability for this technique with test-retest and split-half methods. This approach suggests that traditional concepts of psychometric reliability need not be abandoned in dealing with projective techniques, but that much work remains to be done. Graham (15) studied the reliability of the *Machover Draw-a-Person Test* with a group of 28 graduate students. Following the initial testing he gave them a two-hour lecture on the psychology of human figure drawings. The subjects were then asked to redraw their figures. Little or no changes were found in the second drawing, and a rho of .71 was reported between the two administrations. Arnold and Walter (1) studied the *Rotter Incomplete Sentences Test* responses of 120 freshman college women. With a one-week interval between test administrations, a test-retest reliability coefficient of .82 was reported.

It would seem from the studies reviewed above that satisfactory levels of validity and reliability for projective techniques have not yet been established. Some studies reported coefficients that began to approach the levels necessary for group use, but no reports presented evidence that these tools are sufficiently valid or reliable for individual use.

Normative Procedures

During the period under review there seemed to be a growing interest in securing normative data about children from differing social class levels, subcultural, sex, and age groups. Except for the study by Rohrer and others (37), little was offered in the way of increasing normative data about normals. An attempt was made by Wertheimer (50) to explore the relationships between sociometric data used as a criterion of social adjustment and the *Rorschach*. She classified 200 tenth-, eleventh-, and

twelfth-grade students by sex, IQ, and socioeconomic level and reported *Rorschach* norms for these groups. No significant relationships were found between *Rorschach* indications of social adjustment and sociometric ratings of social adequacy.

Setze and others (40) provided age norms for 216 children between the ages of six and eight who were given the *Rorschach* preceded by a trial card. They found their norms quite similar to those published by Ames. Means, medians, SD's, and percentile ranks for *Rorschach* scoring variables of 50 noninstitutionalized, normal, aged men and women were presented by Light and Amick (28). Fiedler and Stone (13) gave the *Rorschach* to 33 children from low socioeconomic backgrounds and stated that *Rorschach* norms for well-defined samples of children from various socioeconomic groups were needed. They found clear-cut differences in the performance of the children in their sample when they were compared with the Ames's norm group. Rohrer and others (37) published norms of group *Rorschach* performances for 1000 Marines and 374 Naval officers who were unusually well described with respect to sampling characteristics. They reported percentile ranks for each of 39 scoring categories. McCary (32) studied the *Rosenzweig Picture-Frustration* scores of 631 Northern and Southern whites and Negroes, male and female, aged 14 to 22 years, and concluded that there were important performance differences between the pairs of each of these groups. He called for more adequate regional, sex, and racial norms. Reznikoff and Reznikoff (36) categorized 100 second-grade children by sex, race, and socioeconomic status; the greatest performance differences on the socioeconomic variable were found on a family drawing test. Interesting as the above studies of various cultural subgroups may be, the need for adequate norms for normals on all projective devices still exists.

Applications of Projective Techniques

Studies reported in this section do not differ greatly from those reported in the section on validity since studies of applications of projective techniques are generally used as if these instruments were valid. Chahbazi (5) combined a picture and auditory projective technique with the *Ohio State Psychological Test* and an achievement test which were administered to 813 undergraduate agricultural students in an attempt to predict their first-semester grades. He reported an R of .60 with the combined battery as compared to .51 without the projective tests. Wysocki (51) tested 132 men and 85 women with the group *Rorschach*. He used Raven's *Matrices* and *S. P. Test 15* (a British Army verbal test) to estimate intelligence. His findings suggested little relationship between measures of intelligence and *Rorschach* categories; his reported coefficients ranged from .05 to .45.

Three studies reported use of projective techniques in predicting teacher success. Shapiro, Biber, and Minuchin (41) tried out a new cartoon

picture test with 65 student teachers. Their report indicated interscorer reliability coefficients of .47 to .92 but did not present validity data. Kimler (21) found that *Rotter Incomplete Sentence Test* scores and modification of Alexander's *Adult Child Interaction Test* did not predict *Minnesota Teacher Attitude* scores of 58 female practice teachers. He found *Rotter* scores more predictive of data on an unstructured *Behavior Description Blank* when attempts were made to assess the interpersonal relation of his subjects to children in the classroom. An interesting discussion of the relationship of *Rorschach* findings and personality traits described as characteristic of superior teachers was presented by Symonds and Dudek (47), but since only 17 subjects were used, the reported *R* of .60 may be questioned. Light and Amick (28) examined *Rorschach* responses of normal aged.

Mindess (34) attempted to select student nurses. He tested 68 students with the *Wechsler-Bellevue* and the *Rorschach*, scoring the latter with the *Prognostic Rating Scale*. *Wechsler* and *PRS* scores were independent (r=.10) and a multiple *R* of .59 was obtained in prediction of the total nursing grade. The *PRS* score alone correlated .41 with the training grade. The investigator described the sample and the criteria of success more adequately than is done in many such attempts. None of the hypotheses underlying the use of the *House-Tree-Person Test* was supported in a study by Wawrzaszek and others (48) who tested 41 children with severe physical handicaps and compared their scores with a control group which was selected on the basis of age, sex, and IQ. A lighter note was achieved by Meltzoff and Litwin (33) who presented *Rorschach* cards III and VII together with Spike Jones' "Laughter" record to 68 college students who were told not to smile. Significantly more human movement responses were found among the subjects who were successful in inhibiting laughter.

Conclusions

Comparison of the literature on projective techniques covered in the REVIEW by the authors (39) six years ago and that covered during the period of the past three years makes them wonder whether there has been any progress. At that time it was stated that research in this field was needed to separate what could be demonstrated from what was claimed. This distinction is still the central need in research on projective techniques. Korner (24) listed four major problems that seemed significant: (a) problems of the scorer and his interpretations, (b) problems raised by gross qualitative differentiations, (c) problems raised by focus on abnormal rather than normal functioning, and (d) problems raised by confusions that arise from attempts to generalize about real behavior from test behavior. One newer trend has been that of relating projective theory to perceptual process research. Wertheimer (49) discussed this with

particular reference to the *Rorschach* and called for integration of the findings of perceptual research with that of projective research. Baughman (2) also attempted to clarify the perceptual basis of the *Rorschach* technique. In a very thorough review and evaluation of studies carried out with this instrument he noted the increasing need for concern with the development of a unified behavior theory which could encompass all behavior whether in real life or in response to ink blots. He stated that this cannot be done without careful study of stimulus materials. He presented a careful review of *Rorschach* studies on the influence of color, shading, physiological correlates, order, and symbolic meaning, and concluded that the definition of *Rorschach* stimulus effects is more of a task for the future than an accomplishment of the past.

The present reviewers agree wholeheartedly with this statement. It would seem that the time has come to unify the present piecemeal research and to concentrate on a co-operative effort at deriving meaning from the morass of casual empiricism that currently typifies much of the research with projective techniques. Perhaps some group, such as the Society for Projective Techniques, could encourage and manage a major effort similar to that seen in the case of some newer achievement tests where 50 or more experts are called in to produce a major work. Preferable to repeated attempts by single investigators who try to develop new techniques standardized on very small groups and presented with inadequate validation, might be a moratorium until concerted effort is made to demonstrate what current materials can accomplish. This would mean the establishment of adequate normative tables for normals as well as various diagnostic groups, long-term prediction studies, and the elimination of time-worn esoteric uses of these instruments.

Bibliography

1. ARNOLD, FRANK C., and WALTER, VERNE A. "The Relationship Between a Self- and Other-Reference Sentence Completion Test." *Journal of Counseling Psychology* 4: 65-70; Spring 1957.
2. BAUGHMAN, E. EARL. "The Role of Stimulus in Rorschach Response." *Psychological Bulletin* 55: 121-47; May 1958.
3. BOYD, NANCY A., and MANDLER, GEORGE. "Children's Responses to Human and Animal Stories and Pictures." *Journal of Consulting Psychology* 19: 367-71; October 1955.
4. CARTWRIGHT, ROSILAND D. "Predicting Responses to Client-Centered Therapy with the Rorschach PR Scale." *Journal of Counseling Psychology* 5: 11-17; Spring 1958.
5. CHAHBAZI, PARVIZ. "Use of Projective Tests in Predicting College Achievement." *Educational and Psychological Measurement* 16: 538-42; Winter 1956.
6. CHAMBERS, GUINEVERE S., and HAMLIN, ROY M. "The Validity of Judgments Based on 'Blind' Rorschach Records." *Journal of Consulting Psychology* 21: 105-109; April 1957.
7. CHURCHILL, RUTH, and CRANDALL, VAUGHN J. "The Reliability and Validity of the Rotter Incomplete Sentences Test." *Journal of Consulting Psychology* 19: 345-50; October 1955.
8. CRONBACH, LEE J., and MEEHL, PAUL E. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52: 281-302; July 1955.

9. DATEL, WILLIAM E., and GENGERELLI, JOHN A. "Reliability of Rorschach Interpretations." *Journal of Projective Techniques* 19: 372-81; December 1955.
10. DAVIDS, ANTHONY, and OTHERS. "Projection, Self-Evaluation, and Clinical Evaluation of Aggression." *Journal of Consulting Psychology* 19: 437-40; December 1955.
11. EPSTEIN, SEYMOUR; NELSON, JANE V.; and TANOFSKY, ROBERT. "Responses to Inkblots as Measures of Individual Differences." *Journal of Consulting Psychology* 21: 211-15; June 1957.
12. ESCHENBACH, ARTHUR E., and BORGATTA, EDGAR F. "Testing Behavior Hypotheses with the Rorschach: An Exploration in Validation." *Journal of Consulting Psychology* 19: 267-73; August 1955.
13. FIEDLER, MIRIAM F., and STONE, L. JOSEPH. "The Rorschachs of Selected Groups of Children in Comparison with Published Norms: I. The Effect of Mild Hearing Defects on Rorschach Performance. II. The Effect of Socio-Economic Status on Rorschach Performance." *Journal of Projective Techniques* 20: 273-79; September 1956.
14. FINE, REUBEN. "A Scoring Scheme for the TAT and Other Verbal Projective Techniques." *Journal of Projective Techniques* 19: 306-309; September 1955.
15. GRAHAM, STANLEY R. "A Study of Reliability in Human Figure Drawings." *Journal of Projective Techniques* 20: 385-86; December 1956.
16. GRANICK, SAMUEL, and SCHEFLEN, NORMA A. "Approaches to Reliability of Projective Tests with Special Reference to the Blacky Pictures Test." *Journal of Consulting Psychology* 22: 137-41; April 1958.
17. GRIFFIN, DOROTHY P. "Movement Responses and Creativity." *Journal of Consulting Psychology* 22: 134-36; April 1958.
18. HAFNER, ADOLPH J. "Response Time and Rorschach Behavior." *Journal of Clinical Psychology* 14: 154-55; April 1958.
19. JENSEN, ARTHUR R. *Aggression in Fantasy and Overt Behavior.* Psychological Monographs, No. 445. Washington, D. C.: American Psychological Association, 1957. 13 p.
20. JONES, RICHARD M. "The Negation TAT: A Projective Method for Eliciting Repressed Thought Content." *Journal of Projective Techniques* 20: 297-303; September 1956.
21. KIMLER, STEPHEN J. *The Relationship of Parental Identification to the Quality of Interpersonal Relations in the Classroom.* Doctor's thesis. Tempe: Arizona State College, 1958. 217 p. (Typewritten.)
22. KLOPFER, BRUNO, and OTHERS. *Developments in the Rorschach Technique.* Vol. II: *Fields of Application.* Yonkers-on-Hudson, N. Y.: World Book Co., 1956. 828 p.
23. KNOPF, IRWIN J. "Rorschach Summary Scores in Differential Diagnosis." *Journal of Consulting Psychology* 20: 99-104; April 1956.
24. KORNER, ANNELIESE A. "Limitations of Projective Techniques: Apparent and Real." *Journal of Projective Techniques* 20: 42-47; March 1956.
25. LEBO, DELL, and HARRIGAN, MARGARET. "Visual and Verbal Presentation of TAT Stimuli." *Journal of Consulting Psychology* 21: 339-42; August 1957.
26. LEVINE, MURRAY; GLASS, HARVEY; and MELTZOFF, JULIAN. "The Inhibition Process, Rorschach Human Movement Responses, and Intelligence." *Journal of Consulting Psychology* 21: 41-45; February 1957.
27. LEVITT, EUGENE E. "Alleged Rorschach Anxiety Indices on Children." *Journal of Projective Techniques* 21: 261-64; September 1957.
28. LIGHT, BERNARD H., and AMICK, JEAN H. "Rorschach Responses of Normal Aged." *Journal of Projective Techniques* 20: 185-95; June 1956.
29. LINDZEY, GARDNER. "Thematic Apperception Test: The Strategy of Research." *Journal of Projective Techniques* 22: 173-80; June 1958.
30. LINDZEY, GARDNER, and TEJESSY, CHARLOTTE. "Thematic Apperception Test: Indices of Aggression in Relation to Measures of Overt and Covert Behavior." *American Journal of Orthopsychiatry* 26: 567-76; July 1956.
31. LISANSKY, EDITH S. "The Inter-Examiner Reliability of the Rorschach Test." *Journal of Projective Techniques* 20: 310-17; September 1956.
32. MCCARY, JAMES L. "Picture-Frustration Study Normative Data for Some Cultural and Racial Groups." *Journal of Clinical Psychology* 12: 194-95; April 1956.

33. MELTZOFF, JULIAN, and LITWIN, DOROTHY. "Affective Control and Rorschach Human Movement Responses." *Journal of Consulting Psychology* 20: 463-65; December 1956.
34. MINDESS, HARVEY. "Psychological Indices in the Selection of Student Nurses." *Journal of Projective Techniques* 21: 37-39; March 1957.
35. MORE, DOUGLAS M. "The Congruence of Projective Instruments in Personnel Assessment." *Journal of Applied Psychology* 41: 137-40; June 1957.
36. REZNIKOFF, MARVIN, and REZNIKOFF, HELGA R. "The Family Drawing Test: A Comparative Study of Children's Drawings." *Journal of Clinical Psychology* 12: 167-69; April 1956.
37. ROHRER, JOHN H., and OTHERS. *The Group-Administered Rorschach as a Research Instrument: Reliability and Norms.* Psychological Monographs, No. 393. Washington, D. C.: American Psychological Association, 1955. 13 p.
38. ROTHNEY, JOHN W. M., and HEIMANN, ROBERT A. "Development and Applications of Projective Techniques." *Review of Educational Research* 26: 56-71; February 1956.
39. ROTHNEY, JOHN W. M., and HEIMANN, ROBERT A. "Development and Applications of Projective Tests of Personality." *Review of Educational Research* 23: 70-84; February 1953.
40. SETZE, LEONARD A., and OTHERS. "A Rorschach Experiment with Six, Seven, and Eight Year Old Children." *Journal of Projective Techniques* 21: 166-71; June 1957.
41. SHAPIRO, EDNA; BIBER, BARBARA; and MINUCHIN, PATRICIA. "The Cartoon Situations Test: A Semi-Structured Technique for Assessing Aspects of Personality Pertinent to the Teaching Process." *Journal of Projective Techniques* 21: 172-84; June 1957.
42. SHATIN, LEO. "Relationships Between the Rorschach Test and the Thematic Apperception Test." *Journal of Projective Techniques* 19: 317-31; September 1955.
43. SHAW, MERVILLE C., and CRUICKSHANK, WILLIAM M. "The Rorschach Performance of Epileptic Children." *Journal of Consulting Psychology* 21: 422-24; October 1957.
44. SILVERSTEIN, ARTHUR B., and ROBINSON, HARVEY A. "The Representation of Orthopedic Disability in Children's Figure Drawings." *Journal of Consulting Psychology* 20: 333-41; October 1956.
45. SIPPRELLE, CARL N., and SWENSEN, CLIFFORD H., JR. "Relationship of Sexual Adjustment to Certain Sexual Characteristics of Human Figure Drawings." *Journal of Consulting Psychology* 20: 197-98; June 1956.
46. SOMMER, ROBERT, and SOMMER, DOROTHY T. "Assaultiveness and Two Types of Rorschach Color Responses." *Journal of Consulting Psychology* 22: 57-62; February 1958.
47. SYMONDS, PERCIVAL M., and DUDEK, STEPHANIE. "Use of the Rorschach in the Diagnosis of Teacher Effectiveness." *Journal of Projective Techniques* 20: 227-34; June 1956.
48. WAWRZASZEK, FRANK, and OTHERS. "A Comparison of H-T-P Responses of Handicapped and Non-handicapped Children." *Journal of Clinical Psychology* 14: 160-62; April 1958.
49. WERTHEIMER, MICHAEL. "Perception and the Rorschach." *Journal of Projective Techniques* 21: 209-16; June 1957.
50. WERTHEIMER, RITA R. "Relationships Between Specific Rorschach Variables and Sociometric Data." *Journal of Projective Techniques* 21: 94-97; March 1957.
51. WYSOCKI, BOLESLAW A. "Assessment of Intelligence Level by the Rorschach Test as Compared with Objective Tests." *Journal of Educational Psychology* 48: 113-17; February 1957.
52. WYSOCKI, BOLESLAW A. *Rorschach Card Preferences as a Diagnostic Aid.* Psychological Monographs, No. 413. Washington, D. C.: American Psychological Association, 1956. 16 p.

CHAPTER VII

Developments and Applications in the Area of Construct Validity

CHERRY ANN CLARK

THIS issue of the REVIEW marks the first time that an entire chapter has been devoted to construct validity. This presentation, therefore, includes not only a review of recent pertinent publications, but also a résumé of the historical development of the concept with special attention to the influences of the philosophy of science and of theoretical psychology upon the psychometric concept of validity. Also treated are current trends in the *rapprochement* of the theory and methodology of experimental psychology and psychometrics; recent systematic and empirical developments in test construction pertinent to construct validity; and, in conclusion, an evaluation of the apparent heuristic values of the term and the method.

Construct validity departs from classical notions of test validity in that it does not confine itself to the assessment of the extent to which a test score measures an outside criterion. Construct validity is concerned with the logical and empirical investigation of what psychological qualities a test measures (5, 6, 8, 38, 89). It is based upon the experimental evaluation of the behaviorally relevant aspects of a theory. Construct validity may be considered a special case of the general scientific methodology for giving inductive support to the hypothetical regions of a theoretical network (11, 12, 13, 14, 24, 94).

Development of Construct Validity

Construct validity as a method and goal of test construction and evaluation was first proposed in 1954 by the Joint Committee on Test Standards (6). *Technical Recommendations for Psychological Tests and Diagnostic Techniques* emphasized the validation processes and requirements in test development to an unprecedented degree compared even with such recent publications as Gulliksen's *Theory of Mental Tests* (64). Test validation was thought of as involving four different types of judgments and aims; namely, predictive, concurrent, content, and construct validity (6). In predictive and concurrent validity, test performance was considered in relation to how it measured future or present approximation to the criterion as in predicting college success or measuring job performance; in content validity, performance was evaluated in relation to how well it sampled the universe of test items as in a vocabulary test. In construct validity, on the other hand, frequently no definitive criterion was specified; the purpose was conceived as an attempt to clarify at least some of

the measurable characteristics of a trait or function such as intellectual capacity, originality, ego strength, personality structure, or motivation.

The Committee's recognition of the inadequacy of the classical notions of validity in terms of a criterion had been foreshadowed in the preceding decade by several noteworthy publications (7, 34, 55, 57). The dilemma of variations of validity coefficients obtained in repeated test samplings led to the suggestion that the nature of the trait involved be explored rationally and experimentally (34, 44, 55, 57). There was also concern with delimiting sources of error and distortion in the testing process which obstructed efforts to obtain relatively pure measures of traits (7, 57).

Cronbach, for instance, developed the notion of logical validity, based upon the use of deductive and inductive methods of logical analysis to determine the psychological processes that affect test scores (34). He (35, 36) emphasized that subjects' test-taking attitudes as well as the formal character of test questions were important determinates of test responses. His ideas gave impetus to continuing and widespread concern (3, 6, 21, 32, 50, 52, 76, 77, 80, 81, 82, 88, 118, 126, 127).

Guilford developed the concept of factorial validity from his work in factor analysis (56, 57, 59, 84). Factorial validity refers to the amount of the loading or saturation of a test in a given factor; that is, its degree of correlation with a factor. The square root of the communality of a test —the extent to which it measures factors common to other measures— he called relevant validity (58). An outgrowth of this approach was the attempt to define the behavioral characteristics of the factors and the interrelationships of the factorial components of such areas of behavior as reasoning and intellect (61), psychomotor abilities (60), and temperament (62), using a series of interwoven hypotheses (56).

Gulliksen (63, 64) stressed the importance of searching for fundamental and lasting validities contrasted with those that are fortuitous and transient; he used the term *intrinsic validity* to define the process and goal. He outlined a testing program to investigate the intrinsic content validity of achievement tests and the intrinsic correlational validity of aptitude tests.

Goodenough (55) summarized the contributions of tests to the various areas of psychology and described the mental test as an interdisciplinary tool of research rather than a mere instrument to expedite practical measurements. She differentiated test scores as signs and samples (or measurements) of the characteristic under investigation, thereby focusing attention upon the importance of formulating explicit assumptions about the relation of test behavior and scores. She also discussed the need for integrated experimental and psychometric research.

Among others who foreshadowed the altered approach to validity were Jenkins (72) in his criticism of the use of an imperfect criterion, Mosier (104) with his papers on "face validity" and "validity generalization,"

and Anastasi (7) in her statement about the processes involved in the interpretation of test scores.

Expansion in the meaning of, and the reference points for, test validity were but part of the trend in psychometrics which culminated in the emergence of construct validity. There was also a growing concern with testing behavioral hypotheses in experimentally controlled situations. Goodenough's recommendations in 1949 (55) were followed in 1951 by two stimulating articles. Flanagan (44) suggested that test development would be well served if test developers were to recognize the logical relationship between behavior and test measures. He pointed out that tests measure behavior only to the degree that test scores reflect and approximate realistic segments of behavior. He proposed that comprehensive rationales be formulated as a basis for constructing a test, rather than relying on intuitive and practical considerations. Such rationales were to include a description of the behavior to be tested in as many behavioral ramifications as could be conceived, an analysis of the behavior with special emphasis upon the various inferences that could be expected to be made from test performance to behavior, and the precise formulation of item specifications deduced from the analysis of behavior. Travers (125) in a somewhat similar vein discussed the advantages and disadvantages of the rational and technical approaches to test construction. As long as test development was confined to the production of useful instruments to measure a particular thing in a particular situation, he saw little hope of overcoming the stalemate of the last 20 years; he concluded that greater concern with the rational approach would undoubtedly resolve many of the recurring problems facing the test developer.

Advances in statistical methods, especially in multivariate and nonparametric methods, were a necessary adjunct to the procedures required for construct validity (99, 102). In 1950, Eysenck (40) showed how factor analysis could be combined with logical and experimental procedures to investigate complex forms of behavior. He argued that statistical procedures could be incorporated with scientific methodology to elucidate some of the taxonomic phases of psychological investigation.

Peak's "Problems of Objective Observation" (107) was one of the most significant precursors to the 1954 statement on construct validity. Her presentation has remained one of the outstanding introductions to the complex aspects of theory construction and confirmation to be found in psychological literature. She gave a well-reasoned triple classification of test validity: face validity, validity with prediction to a criterion for some particular purpose, and validity involving testing predictions from a theory. She used the term *functional unities* to designate the dynamic constructs which are the substance of psychological research. She described clearly the methodology involved in constructing a theory from observationally limited hypotheses and subjecting them to experimental verification. She discussed the bearing of the experimental manipulations

needed to demonstrate concomitant variation, interdependence and/or dependence of events on such psychometric methods as item analysis, scaling, intertest correlation, and factor analysis and other multivariate methods. She indicated ways in which experimental and psychometric methods could be combined to investigate such complex functions as hostility when such behaviors are embedded in an explicit, testable theoretical structure. She also included a critique of the use of models from other sciences in handling the theoretical formulation of psychological phenomena.

Bindra and Scheier (16) in a short but cogent article recommending the combination of experimental variation with psychometric variables, contributed to the trend away from the unalloyed empiricism dominating test development since Binet. They compared the characteristics of such psychometric variables or constructs as Murray's needs, Allport's traits, and Cattell's factors (derived from the empirical methods used in the study of individual differences) with the more theoretically oriented variables of personality research which emerged from Lewin's and MacKinnon's studies. They showed how systematic variation of the experimental variables could help specify sources of variation and error in test construction and evaluation. These sources of error have not been clarified by ordinary psychometric procedures; indeed, in the opinion of Bindra and Scheier, psychometric research has not been able to organize a program which could adequately conceptualize the variables it has worked with. They suggested that the combined methodology would bring into relief the relationship between variant and invariant aspects of personality.

Butler published a thought-provoking article in 1954 (20) criticizing the failure of psychometrists to incorporate psychological theory into test theory. He commented that mental test theory has consisted of a set of postulates and theorems more or less rigorously applied to test construction regardless of the content or purpose of the test. Such formalism has forestalled the psychometric exploration and clarification of significant behavioral variables. Most psychometric devices, including personality inventories, have consisted of a set of items subjectively selected by the test designer without particular concern for the relation of the items to psychological theory. Even introversion-extroversion scales based upon Jungian typology have been only remotely associated with Jung's personality theory.

Butler demonstrated how a psychological theory and model such as Tolman's could be used to derive a series of hypotheses about the characteristics of inventory items. He discussed how such hypotheses could be integrated into a useful experimental framework of intervening or independent variables. He reasoned that according to Tolman's theory the personality inventory items most promising for predicting behavior would refer to behavior space, feeling states, and the belief-value matrix; least valuable items would refer to the need systems, for which no direct

behavioral self-report representation has been found. He mentioned the use of Stephenson's Q-methodology to study the personal value system and the need for developing a metric system suitable to the content and assumptions of the theoretical model.

The preliminary proposal of the American Psychological Association Committee on Test Standards (5) used the term *congruent validity* to indicate the correspondence between scores on a test and other indications of a psychological state or attribute. The procedure is exemplified in a study by Abernethy and White (1) which validated the *Guilford GAMIN Test* with laboratory measures of vigor and motility and other behavior signs interpreted as indicating dominance and leadership. Such validation was intended for tests measuring a construct arising from some theory, and was considered conterminous with the evaluation of a theory itself. Suggestions were given for developing a testing program which would implement the search for significant psychological traits.

The 1954 "Technical Recommendations for Psychological Tests and Diagnostic Techniques" (6) replaced the term *congruent validity* with *construct validity* and gave more extensive treatment to the logical and empirical bases for the validational process. Not only was construct validity related to the development of a testable psychological theory, but also to the other types of validity, showing the interdependence of the validating procedures and results. It was emphasized that the numerical measures of behavior obtained in the process of construct validation could not be interpreted as validity coefficients, but rather as partial evidence in the process of clarifying the logical and measurable characteristics of an attribute. Numerical indexes were viewed merely as facets of the hypothetical structure supporting a theory. It was considered essential that test developers specify what aspects of a theory had been subjected to validational research.

Construct validity was further discussed in terms of the construction and technical evaluation of such instruments as ability tests, personality and interest inventories, and projective and related clinical methods, each measuring device having its own peculiar requirements for theoretical and empirical testing.

The distinction between the *behavior-equivalence* and the *behavior-relevance* of a construct was outlined against a theoretical and methodological framework. Construct validity was explained as mainly concerned with the behavioral-relevance of test measures, whereas content validity was determined by behaviorally equivalent test items. Because of the incompleteness of current psychological theory and the consequent vagueness in the definition and use of many psychological terms or constructs, the process of construct validity was described as being fraught with intellectual dangers but nonetheless essential to the development of psychological testing.

In 1955, Cronbach and Meehl (38) presented their comprehensive treatment of construct validity. They argued for the need of a concept

of validity which was not confined to validation in terms of specific criteria or immediately observable variables and for the feasibility of validating constructs interrelated with other constructs in a nomological network (94, 108), some of which could be associated with observable behavior, but others only inferred by the prediction of imputed relations. Gradually, constructs initially unanchored by observational definition and measurement could be brought closer to confirmable constructs as the relationships among constructs or variables could be specified. They indicated how a relatively complex system of interrelated constructs could cope with the seemingly unfathomable aspects of behavior found in attempting to assess intellectual ability, social conformity, and the theory of paranoia and aggressiveness (to mention but a few of the areas of behavior which psychological tests have tried to penetrate). They took issue with the operationalism recommended by Spiker and McCandless (119) as being too rigid to deal with the problems at hand and scientifically unnecessary.

They described an experimental methodology combining such statistical designs as correlation and factor analysis, such psychometric methods as content validity, interitem and intertest correlations, and the planned manipulation of variables affecting the test situation and individual and group differences. They gave several examples of possible applications, showing how the investigator could handle both positive and negative experimental data in the process of construct validation.

Jessor and Hammond (74) stressed that the psychometrist is as obliged as the experimentalist to recognize the implications of the theoretical model for designing the tests of his hypotheses. Theory must be used to develop a test just as theory must be used to structure an experiment. They maintained that one of the difficulties in validating the *Taylor Manifest Anxiety Scale* was that the test, intended to measure drive according to the Hullian paradigm, was not devised so that drive could be inferred from test responses. They questioned whether a self-report inventory consisting of items selected on the basis of clinical judges' opinions is capable of measuring drive as conceptualized in Hullian theory; as yet no correspondence between the Hullian construct of drive and the *A-Scale* construct of presumed drive has been found. They recognized that inferences must be made from observable data to a hypothesized construct in any well-planned theoretical test, but cautioned that there must be a link between the evidence and the construct by virtue of the experimental elicitation and variation of the behavior in question, not by mere implication or intuition that certain responses are a measure of a construct. They emphasized, as Peak (107), Butler (20), and Cronbach and Meehl (38) had previously shown, that the crux of testing any theory using psychometric procedures is that the theory must have a prior role in guiding test development, not one subsequent to the construction of the test.

They pointed out that since there was a lack of explicit relationship between *A-Scale* items and drive properties as defined in Hull's theory, even Taylor's (19, 26, 30, 69, 71, 81, 91, 128) could not overcome the limitations of the scale imposed by the assumptions that anxiety or emotionality can be assessed by paper-pencil self-report about nonverbal responses. They pointed out that items, such as those in the *A-Scale*, are excessively vulnerable to the response set distortions of social desirability, deception, and self-insight. They felt that it would have been better to attempt an evaluation of these factors as the scale was constructed rather than to reason after the fact concerning the influence these response sets might have in determining responses. They mentioned other shortcomings of the scale, including the dichotomous scale approach to measuring intensity of response.

Jessor and Hammond (74) gave comparatively more attention than other writers on construct validity to depicting the logical and methodological necessity for explicit alternative hypotheses in the process of confirming any set of hypotheses. They showed that the derivation of the *A-Scale* ignored the possible alternative implications of the relation of anxiety to drive as well as other reasonable interpretations of what the scale measured such as intellectual function, motivational status, and psychopathological involvement. In addition, they indicated how the nomological network surrounding such a construct as drive could be elaborated to investigate the diverse properties of the construct and how conditional definitions of the construct form the logical framework for relevant experimental manipulation of the variables. They insisted that the designation of the variables must be made independently of the tests of the relationship implied in the construct under investigation.

Loevinger in her long and rather technical monograph (89) concentrated on the psychometric scheme necessary for construct validity. Her rationale was developed for the limiting case of objective tests based upon the dichotomous scoring of items having a determinable difficulty level; nevertheless, it supplied an interesting and provocative model from which others might take inspiration for grappling with the extremely difficult problems of scale construction, item constitution, and scoring.

She asserted that the most fruitful development in psychometric devices will be found in the measurement of traits which have real existence in some sense. She likened a trait to a parameter: It was what psychologists have tried to understand; a construct, like a statistic, was the current best estimate of the trait. She defined the elements of construct validity as the test, the traits measured, and what the tester asserted he has measured; in other words, construct, interpretation, and theory. The degree of internal structure of the items and the magnitude of external correlations she conceived as constituting the psychometric evidence for construct validity; the nature of the structure, the content of the items, and the nature of the external relations for her constituted the psychological evidence.

Test behavior, consisting of responses to items, she reasoned, must be treated as both signs and samples. Because they represented samples of behavior, they were assumed to be subject to the same laws as behavior in general; because they represented signs, inferences could be drawn from the patterning of test responses to the organization of other behavior.

She posited that construct validity had three aspects: the substantive component, the structural component, and the external component. The substantive component, which for her included such concepts as homogeneity, functional unity, and content validity, was the guiding consideration in selecting items. Items to be useful in establishing the substantive validity of a test had to be drawn from a very broadly defined pool or region of behavior and had to be chosen so that all possible aspects of the inferred trait, including alternative theories of the trait, were represented (in accordance with Brunswik's model of representative design). Substantive validity was defined as the extent to which the content of the items included in and excluded from the test could be accounted for in terms of the putative trait and the context of measurement. She devoted considerable space to the discussion of the structural component and the concept of structural validity. Briefly, she defined the structural component of validity as the extent to which structural relations between test items parallel the structural relations of other manifestations of the trait being measured. In other words, structural validity was referred to the similitude of the structural model and the structural characteristics of nontest evidence of the trait as well as to the degree of interitem structure. The structural component could be implemented by using different available quantitative models such as various scaling procedures, class models as in certain clinical and personality-theory assumptions about behavior, and dynamic models, still inadequately conceptualized but crucial to personality investigation. She discussed the possible contributions of pattern analysis and configural scoring to structural analysis and the difficulties involved in selecting appropriate structural models to use in a program of construct validation. The external component and external validity she presented in the light of the relationship of a test score to the extratest evidence of the imputed trait. She mentioned such procedures as factorial patterning and suppressor variables as problems to be considered in evaluating the nontest evidence for the trait.

She argued that the standard reliability formulas were inappropriate in construct validity psychometrics, for the assumption that the errors of measurement in two parallel tests were equal to zero did not hold; each test administration had to be considered as influenced by "secular trends," the changes in behavior over time. Such matters, she advocated, must be given a legitimate place in test theory if the problems of trait consistency, developmental changes, intra-individual differences and pathological and situational processes were to be handled adequately.

She summarized her program for test construction by pointing out that all possible sources of data, psychometric and behavioral, must be assembled in convergent operations to define the trait being explored. Test scores had to be overdetermined if they were to be acceptable in trait validation; all possible theoretical alternatives were to be examined before confirmation of any trait could be asserted.

Empirical Studies on Construct Validity

This section does not represent an exhaustive listing of recent publications involving the methods and aims of construct validation. The reviewer has limited the presentation to epitomizing current trends in published articles relevant to or purportedly constituting construct validation.

Few have been the subjects subsumed under the headings of individual differences and personality and clinical assessment which have not at least been stated in terms of a construct validation goal. Recent reviews of these topics (4, 26, 30, 33, 48, 50, 51, 71, 73, 79, 81, 82, 89, 90, 95, 115, 124, 128) frequently evaluated the achievements in test development by the degree to which research conformed to the recommendations for construct validation. Some research labeled construct validation did not make so significant a contribution to the goals of construct validity as research which emphasized the development of a theoretical or methodological structure in some particular area.

The *Taylor Scale of Manifest Anxiety* was one of the early efforts to apply a personality test to the assessment of behavior in terms of a theoretical model (71, 73, 82, 114). In spite of the persistent efforts to investigate the construct "anxiety" as a drive in combined experimental and psychometric research, the result and the interpretations of what the test measured were contradictory. There was much adverse criticism about such aspects of the test as item selection, the definition of the concept, and the slowness in modifying the stated theoretical structure and inferences in view of experimental research.

Siegman (117) remarked that his study comparing various measures of anxiety with scores from the *MAS* suggested that the *MAS* did not have significant criterion validity but had considerable construct validity. This study, like others cited in the review (23, 71, 78, 80, 91, 114, 117, 118) fell short of the methods and aims envisioned by writers concerned with improving psychometric methodology (16, 20, 89, 107, 125). The relationships among the variables used as measures of anxiety were not clearly stated in reference to a rational structure external to the test situation. The definitions of the measures were stated for the specific operations involved in the test situation without shedding any light upon the more abstract and general character of the construct "anxiety" (91, 114, 128). Siegman's work, however, probably represented a necessary

step toward evolving an adequate theoretical and empirical outline for the eventual clarification of the psychological processes imputed to the construct.

Various other pools of items from the *Minnesota Multiphasic Personality Inventory* were subjected to construct validation. *Barron's Ego Strength Scale,* for example, was investigated. Tamkin (121) and later Tamkin and Klett (122) obtained dubious results about the worth of *Barron's Ego Strength Scale* for determining ego strength. They used psychiatric diagnoses, *F*+ percent, the *Pascal-Suttell Critical Item,* the *F Scale* of the *MMPI,* and *Wechsler-Bellevue Intelligence Test* scores as the extra-test behavioral indexes for the imputed trait of ego strength. These studies in the reviewer's opinion are examples of rather makeshift approaches to the complex and exacting problems of construct validation. One of the shortcomings was that the construct of "ego strength" was not defined adequately apart from the operations used to investigate it, thereby introducing circularity in the definition. Also, the hypothetical network to which the construct of "ego strength" reputedly belonged was not clearly stated in the two articles. Certain assumptions were made—for instance, that diagnostic groups represented different levels of ego strength—for which no test was suggested. On the positive side, however, the investigators did record the negative evidence for their hypothesis, a necessary event in any construct validation program. However, negative evidence, it was pointed out (38), is no more compelling than positive evidence in testing a poorly defined concept in a sketchily outlined theoretical structure.

Bernardin and Jessor (15), recognizing the need for careful specification of the components or properties of test constructs and the relationship between test and theory, undertook to investigate dependency. They attempted to associate the psychometric personality test behavior called dependency to experimentally induced behavior. *Dependency* they defined operationally as including reliance on others for approval or the importance of approval from others, reliance on others for help or assistance, and conformity to opinions and demands of others. They used the two variables of the *Edwards Personal Preference Schedule* called deference and autonomy. They conducted three experiments to test three hypotheses related to behavioral concomitants of dependency under varying experimental conditions. They concluded that subjects scored as dependents in their scheme showed greater reliance on others for approval and help but not necessarily greater group conformity than those classified as independents. Noteworthy about this investigation, in addition to the rather well-defined use of terms and the precisely stated hypotheses and experimental tests, was the critical summary with which Bernardin and Jessor concluded their paper. They succinctly enumerated the deficiencies in their experimental design, the inadequacies in the psychometric procedures, possible effects on the results caused by uncontrolled

and confounding conditions, and proposed modifications in future research to provide evidence for the construct of "dependency" as they defined it. This study, like the one by Liverant (87), was among recent publications which most nearly approximated what has been recommended as adequate research toward construct validation.

Silverman (118) also examined the *Edwards Personal Preference Schedule* for construct validity, but the study was not so comprehensive or well executed as the one by Bernardin and Jessor. He concluded that the forced-choice questions limited the distortion introduced by social desirability.

Liverant (87) went even farther than Bernardin and Jessor toward the goal of developing a research program for investigating the behavioral referrents of a conceptual system. Bernardin and Jessor explored only a relatively limited segment of behavior with but a few allusions to the systematic ramifications of the construct under study. Liverant, on the other hand, started with a larger theoretical framework. He adapted Rotter's social learning theory to the task of constructing a test to measure a hypothetical set of interrelated needs. He attempted to select the test items used to elicit responses pertinent to the measurement of needs in a manner dictated by the hypothetical network deduced from the theory. He tried to overcome the problems of social desirability (118) in response sets common to personality inventories by devising a series of forced-choice items. In his conclusions he noted that the use of judges to select from the pool of available items those which could be matched as parallel and opposite indexes of the needs he had hypothesized would have to be modified in future studies; the judges inevitably introduced their idiosyncratic needs into item selection. Sex differences emerged in the results which were not accounted for by the prior theoretical formulations. Liverant indicated that the theory would have to be re-evaluated to formulate alternative hypotheses which in future investigations would permit better prediction.

In constructing a psychometric test of his hypothetical need system, he postulated that he could infer need by predicting the behavior that would follow particular kinds of events and reinforcements. Certain values, he speculated, would be reinforced by certain external events, whereas others would not. He based his experimental procedure on the expectation of concomitant change in behavior in a specified situation. Psychometrically he derived measures of internal consistency, using the Spearman-Brown formula. Then he established the stability of the measures over time. He used Thurstone's centroid method based on interitem correlations to determine the factorial validity of the various measures of needs. Three factors emerged: social recognition, social love and affection, and academic recognition. The last factor was not so valid as the first two. He compared these factors with those which have emerged from the *Edwards Personal Preference Inventory* studies and suggested that there were similarities in several of the constructs. He cautioned that

his use of test administration instructions to control one of the variables important in his investigation might not have been a sufficient control for the circumstances. In spite of the various limitations in this experiment, it was one of the most conscientiously planned and executed investigations to appear since construct validation has been in vogue.

Campbell and Tyler (23) conducted an experiment to assess the construct validity of work-group morale measures using an experimental design based on convergent analysis (22, 89). Self-descriptions by members of a group were compared with statements about the morale of the individuals by members of other groups who knew the group members. Such a procedure recognized that both measures were fallible and biased. The authors maintained that to the extent to which the biased measures could be assumed to be independent, the two measures validated each other and the construct to which they referred.

This rather ingenious application of part of the method of convergent validation (22) should be subjected to further examination before it becomes widely used.

Jones (75) developed a variation of the *Authoritarian Scales* (31, 65, 83, 109, 111, 112, 120, 123) and explored the construct validity of several of the measures. For the most part he confined his validation procedures to correlations with other known scales. He did, however, present a series of operationally defined behaviors to be explored by the test. He apparently was aiming at the test validation of the interrelationships of behavior defined to some degree independently of the testing situation.

Kaess and Witryol (77) subsequently subjected the *Z Scale* (75) to a form of construct validation by intercorrelating scales from the *California F Scale*, the *Guilford-Zimmerman Temperament Survey*, and the *Allport-Vernon-Lindzey Study of Values Test*. They concluded that there was a significant relationship among the behaviors reflected in the concepts of anxiety, hostility, rigidity, and dependency found in the various tests. It appeared that the forced-choice form of the test tended to eliminate the distortion of the test scores associated with the set to respond in the socially expected way.

Hart (65) investigated the effects of the maternal attitude toward authoritarianism upon child-rearing practices in terms of the construct validity of the procedure. This rather well-planned and circumscribed investigation was typical of the growing concern with including construct validation in a testing procedure. Leblanc (86) also directed her sociological research toward ascertaining the construct validity of several variables. It is hoped that the mere use of the term *construct validity* in the introduction and summary of investigations will not be construed as meeting the requirements of the construct validity approach.

Cattell (28) explored his *16 P.F. Test* for construct validity, using a series of factor analyses. Dahlstrom (39) criticized Cattell's use of the

term *construct validation* on the basis that it was substituted for the measure of internal consistency, a psychometric method useful but not sufficient for construct validation (88). Karson and Pool (78) also examined the construct validity of the *16 P.F. Test* by comparing certain of the factors with *MMPI* profiles. Little of a positive nature can be said for research which does not clearly indicate the limitations for the conclusions offered in the name of construct validity.

Keehn (80) experimented with the repeated use of visual-motor tests to assess the variation in performance from time to time as an experimental variable in construct validity. Secular testing (89) required a great deal more study to determine intro-individual variation over time, especially with projective testing where the distinction between behavior as signs and samples has remained rather poorly conceptualized.

Methodological and Substantive Developments Pertinent to Construct Validity

As mentioned in the foregoing section, much research furthering the process of construct validation was not called by the name. Some of such research was concerned with the systematic definition and organization of a segment of behavior; other research concentrated on methodological procedures; a few studies combined both substantive and methodological contributions.

The area of psychomotor behavior (45, 46, 47, 60) was reviewed and formulated in terms of testable hypotheses. A number of psychomotor functions were investigated by using factor analytic methods combined with variations in experimental conditions (46, 47). Guilford (60) summarized current concepts related to psychomotor abilities in a more or less systematic network of significant heuristic value.

Michael (101) described spatial-visualization abilities in terms of a combined experimental and testing verification of the variables; Michael and others (103) further explored the area.

Bruner, Goodnow, and Austin attempted to develop a theory of thinking, which needed further verification to determine the validity of the concepts and functional relationships included in the theory (82). Kelly's theory about how people acquire and use interpersonal concepts received some experimental investigation using a qualitatively oriented inventory (91).

Blanchard (17) evaluated a diversity of data about intellectual functioning to delineate the conditions and characteristics leading to a function he called intellectual inhibition. Nadelman (106) also described several situational conditions and personality traits found to be important determinants of the quality of conceptual thinking. Marx (97) gave careful attention to defining concepts related to problem-solving behavior and suggested an experimental program rich with heuristic implications.

Lawson and Marx (85) performed a similar service for the field of frustration, analyzing the contradictory evidence from the various theoretical positions.

Several studies were aimed at overcoming the stalemate in research on rigidity (21, 33, 71, 82). Himelstein (70) redefined the concept and used a model suggested by some of his earlier work on Helson's adaptation level theory to test his hypotheses. Ainsworth (2) also suggested that an altered definition of rigidity could surmount the impasse. He related rigidity to behavioral changes under conditions of situational insecurity and stress.

Yates (129) made a very significant contribution to mental testing in his comprehensive review of tests for brain damage. The area was much in need of the construct validation approach to supply objectivity to this important aspect of clinical assessment.

The authoritarian personality was the subject of many empirical investigations (75, 77, 83, 109, 120). Christie and Cook (31) evaluated the evidence relating to the authoritarian personality and specified the nature of the discrepant findings. They discussed the methodology of the studies, made critical recommendations about the future use of the terminology evolved during the past decade, and suggested problems for future exploration. Such ambitions and objective appraisals of the state of affairs in a given area are to be commended.

Titus and Hollander (123) reviewed the use of the *California F Scale* over a five-year period. They brought into focus some of the antecedent conditions which had been found to contribute to test-score variances. Such critiques should invite more systematic investigation of relevant variables, an indispensable aspect of construct validity.

Rokeach (110) amassed evidence from a number of sources in a carefully conceived formulation about dogmatism. He stated his variables and experimental conditions in a Lewinian model. One investigation (49) which used factor analysis gave some indication of how profitable such a formulation could be in guiding experimentation.

Cattell (27) and Cattell and Scheier (29) reviewed a number of tests to classify test behaviors related to anxiety. Guilford and Zimmerman (62), were also instrumental in the taxonomic clarification of several areas of personality. A more extensive treatment of his contributions can be found in Chapter III.

McClelland and others (92) developed a theory around one of Murray's need-press attributes. Much research has accumulated (32, 71, 73, 82) by use of numerous *ad hoc* hypotheses. Several general critiques of the hypotheses tested and the methods used appeared in recent years (124, 128). Doubt was still current about what the *TAT* measured. Papers, such as the ones by Getzels and Walsh (52) and Jones (76), should assist in clarifying the relation of test responses to behavior. Getzels and Walsh, with construct validation as a goal, compared the responses

of subjects to paired projective and objective items to throw light on the propensity of different types of questions to tap various levels of personality defenses.

In the field of intellectual and aptitude testing, Michael (100) outlined a comprehensive rationale for testing high-level personnel, while Ryans (113) listed the problems involved in assessing criteria for teacher selection in terms of the construct validity of the various criteria. Such formulations did not implement the more abstract and complex goals of construct validity, but they gave order to a body of data so that psychometric and practical inferences could be made beyond the level of discrete criteria validations.

Dahlstrom (39) reviewed the contributions of factor analysis to clinical research, emphasizing the assumptions and requirements necessary for appropriate use in developing tests and assessing areas of personality. He discussed the use of Eysenck's criterion analysis (40) to explore the empirical relationships among criteria. He commented that factor analysis did not solve the difficulty of determining the reality of the factors derived; this goal he considered best attained by a construct validity approach.

Falk (41) analyzed the different assumptions underlying nomothetic and idiographic models for personality investigation. He suggested that the two methods were not to be treated as antithetical but rather as corroborative. Other reviews which added materially to the methodology of construct validity included those by Furst and Fricke (50) and Jenkins and Lykken (71).

Loevinger (88) proposed a redefinition of the term *ego development* as a trait, amenable to experimental determination. She hypothesized that ego development could be conceptualized in four testable dimensions: (a) maturity as represented by constantly increasing distance from self ("objective insight"); (b) tendency to increase constantly with age; (c) growth with intelligence, education, and social status; and (d) progressive strengthening with psychotherapy. The implications of her proposed research included the following: self conception and mode of verbalization about self as aspects of ego development, measurement of ego development by verbal means as a complex and challenging procedure, and measurement by nonverbal means as an impossible goal. She urged caution in using such readily available samples as clinical populations in studying ego development. She also emphasized the need for a rational approach to test construction with the aim of arriving at psychologically meaningful goals.

Construct Validity and the Philosophy of Science

The philosophy of science embodied in construct validation had its critics as well as its advocates. Bergmann (11, 12, 13, 14), Brodbeck

(18), and Mosier (105) ably discussed the contributions from the philosophy of science to recent systematic and methodological developments in psychology. Brodbeck specifically related the development and purposes of construct validity to recent trends in the analytical definition of concepts, the relation of concepts to the theories and hypotheses in which they were organized, and the behavioral verification of concepts and dynamic relationships.

The critics of the theoretical formulation of construct validity have not been so outspoken as the advocates; undoubtedly as more research demonstrates the difficulties in implementing programs of construct validation, critics will become more vocal. Bechtoldt (9) made a detailed epistemological analysis of the approach outlined by Cronbach and Meehl (38). He suggested that their formulation of response-defined constructs or attributes created a hiatus between the construct and the multidetermined causes or antecedents of a particular behavior. He felt that their procedures would lead only to circularities or tautologies. He asserted that the use of an empirically based set of concepts oriented to the variables influencing behavior would reduce the frequency of such tautologies. He accepted the plausibility of deriving concepts from observed performance but admonished that concepts had to be defined independently of the performance under investigation.

Additional arguments about the place of concepts in a hypothetico-deductive system occupied much of the attention of theoretical and experimental psychologists (37, 53, 54, 96, 98, 116, 122). Psychometrists were relative late-comers to the field. Among the questions which must be worked on if construct validity is to attain a legitimate position theoretically and empirically, is the degree to which a concept or construct should be allowed to depart from the immediate behavioral observation and to approach an abstract and general specification of inferred behavior (10, 14, 18, 25, 53, 54, 66, 67, 68, 93). Similar problems in the use of various kinds of hypotheses (12, 42, 43) will require intelligent examination. The role of models (14, 18, 20, 43, 107) in psychological testing has not been adequately explored; most testing has been confined to the use of molecular models and statistical rather than nonstatistical laws and methods.

Psychologists, especially test developers, have concerned themselves to only a limited extent with the stimulating, complex, and frequently perplexing issues in the philosophy of science. Only the future will indicate whether psychological testing will benefit from the influences of philosophy of science.

Conclusions

The evaluation of the heuristic impact of the construct validation movement must await future development. In the short space of time since the

concept has been identified, the salutary trends noted earlier in test construction combining psychometric and experimental methods with theoretical formulations continued. Psychological testing became increasingly imbued with the theoretical and methodological sophistication which has characterized experimental psychology for some time. Experimental psychology, in its turn, showed increasing concern for dynamic and general aspects of behavior. As this review has pointed out, the work of construct validation is arduous but rewarding to the test developer interested in being part of the main stream of the philosophy of science and behavioral sciences.

Bibliography

1. ABERNETHY, ETHEL M., and WHITE, JAMES CLYDE, JR. "Correlations of a Self-Inventory of Personality Traits with Laboratory Measures of Vigor and Motility." *Journal of Social Psychology* 40: 185-88; August 1954.
2. AINSWORTH, LEONARD H. "Rigidity, Insecurity, and Stress." *Journal of Abnormal and Social Psychology* 56: 67-74; January 1958.
3. ALPER, THELMA G. "The Interrupted Task Method in Studies of Selected Recall: A Reevaluation of Some Recent Experiments." *Psychological Review* 59: 71-88; January 1952.
4. ALPER, THELMA G. "Predicting the Direction of Selective Recall: Its Relation to Ego Strength and *N* Achievement." *Journal of Abnormal and Social Psychology* 55: 149-65; September 1957.
5. AMERICAN PSYCHOLOGICAL ASSOCIATION, COMMITTEE ON TEST STANDARDS. "Technical Recommendations for Psychological Tests and Diagnostic Techniques: Preliminary Proposal." *American Psychologist* 7: 461-75; August 1952.
6. AMERICAN PSYCHOLOGICAL ASSOCIATION, AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, and NATIONAL COUNCIL ON MEASUREMENTS USED IN EDUCATION, JOINT COMMITTEE ON TEST STANDARDS. "Technical Recommendations for Psychological Tests and Diagnostic Techniques." *Psychological Bulletin* (Supplement) 51: 1-38; March 1954.
7. ANASTASI, ANNE. "The Concept of Validity in the Interpretation of Test Scores." *Educational and Psychological Measurement* 10: 67-78; Spring 1950.
8. ANASTASI, ANNE. *Psychological Testing.* New York: Macmillan Co., 1954. p. 121-51.
9. BECHTOLDT, HAROLD P. "Construct Validity of Construct Validity." Paper presented to the American Psychological Association, September 1958. Symposium on Construct Validity. Iowa City: State University of Iowa, 1958. 16 p. (Mimeo.)
10. BECK, LEWIS W. "Constructions and Inferred Entities." *Philosophy of Science* 17: 74-86; January 1950.
11. BERGMANN, GUSTAV. "The Logic of Psychological Concepts." *Philosophy of Science* 18: 93-110; April 1951.
12. BERGMANN, GUSTAV. *Philosophy of Science.* Madison: University of Wisconsin Press, 1957. 181 p.
13. BERGMANN, GUSTAV. "On Some Methodological Problems of Psychology." *Readings in the Philosophy of Science.* (Edited by Herbert Feigl and May Brodbeck.) New York: Appleton-Century-Crofts, 1953. p. 627-36.
14. BERGMANN, GUSTAV. "Theoretical Psychology." *Annual Review of Psychology.* Vol. 4. Stanford, Calif.: Annual Reviews, 1953. p. 435-58.
15. BERNARDIN, ALFRED C., and JESSOR, RICHARD. "A Construct Validation of the Edwards Preference Schedule with Respect to Dependency." *Journal of Consulting Psychology* 21: 63-67; February 1957.
16. BINDRA, DALBIR, and SCHEIER, IVAN H. "The Relationship Between Psychometric and Experimental Research in Psychology." *American Psychologist* 9: 69-71; February 1954.

17. BLANCHARD, WILLIAM H. "Intellectual Inhibition and the Search for Scientific Truth." *Journal of Social Psychology* 47: 55-70; February 1958.
18. BRODBECK, MAY. "The Philosophy of Science and Educational Research." *Review of Educational Research* 27: 427-40; December 1957.
19. BRONFENBRENNER, URIE. "Personality." *Annual Review of Psychology*. Vol. 4. Stanford, Calif.: Annual Reviews, 1953. p. 157-82.
20. BUTLER, JOHN M. "The Use of a Psychological Model in Personality Testing." *Educational and Psychological Measurement* 14: 77-89; Spring 1954.
21. BUTLER, JOHN M., and FISKE, DONALD W. "Theory and Techniques of Assessment." *Annual Review of Psychology*. Vol. 6. Stanford, Calif.: Annual Reviews, 1955. p. 327-56.
22. CAMPBELL, DONALD T., and FISKE, DONALD W. "The Multitrait-Multimethod Matrix in the Validational Process." Paper Presented to the American Psychological Association, September 1958. Evanston, Ill.: the Authors (Department of Psychology, Northwestern University), 1958. 18 p. (Mimeo.)
23. CAMPBELL, DONALD T., and TYLER, BONNIE B. "The Construct Validity of Work-Group Morale Measures." *Journal of Applied Psychology* 41: 91-92; April 1957.
24. CAMPBELL, NORMAN R. "The Structure of Theories." *Readings in the Philosophy of Science*. (Edited by Herbert Feigl and May Brodbeck.) New York: Appleton-Century-Crofts, 1953. p. 288-308.
25. CARNAP, RUDOLF. "The Methodological Character of Theoretical Concepts." *Minnesota Studies in the Philosophy of Science*. Vol. 1. (Edited by Herbert Feigl and Michael Scriven.) Minneapolis: University of Minnesota Press, 1956. p. 38-76.
26. CARROLL, JOHN B. "Individual Differences." *Annual Review of Psychology*. Vol. 5. Stanford, Calif.: Annual Reviews, 1954. p. 127-48.
27. CATTELL, RAYMOND B. "The Conceptual and Test Distinction of Neuroticism and Anxiety." *Journal of Clinical Psychology* 13: 221-33; July 1957.
28. CATTELL, RAYMOND B. "Validation and Intensification of the Sixteen Personality Factor Questionnaire." *Journal of Clinical Psychology* 12: 205-14; July 1956.
29. CATTELL, RAYMOND B., and SCHEIER, IVAN H. "The Nature of Anxiety: A Review of Thirteen Multivariate Analyses Comprising 814 Variables." *Psychological Reports* 4: 351-88; Monograph Supplement 5, June 1958.
30. CHILD, IRVIN L. "Personality." *Annual Review of Psychology*. Vol. 5. Stanford, Calif.: Annual Reviews, 1954. p. 149-70.
31. CHRISTIE, RICHARD, and COOK, PEGGY. "A Guide to Published Literature Relating to the Authoritarian Personality Through 1956." *Journal of Psychology* 45: 171-99; April 1958.
32. CLARK, RUSSELL A., and MCCLELLAND, DAVID C. "A Factor Analytic Integration of Imaginative and Performance Measures of the Need for Achievement." *Journal of General Psychology* 55: 73-83; July 1956.
33. CRONBACH, LEE J. "Assessment of Individual Differences." *Annual Review of Psychology*. Vol. 7. Stanford, Calif.: Annual Reviews, 1956. p. 173-96.
34. CRONBACH, LEE J. *Essentials of Psychological Testing*. New York: Harper and Brothers, 1949. p. 48-59.
35. CRONBACH, LEE J. "Further Evidence on Response Sets and Test Design." *Educational and Psychological Measurement* 10: 3-31; Spring 1950.
36. CRONBACH, LEE J. "Response Sets and Test Validity." *Educational and Psychological Measurement* 6: 475-94; Winter 1946.
37. CRONBACH, LEE J. "The Two Disciplines of Scientific Psychology." *American Psychologist* 12: 671-84; November 1957.
38. CRONBACH, LEE J., and MEEHL, PAUL E. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52: 281-302; July 1955. Reprint: *Minnesota Studies in the Philosophy of Science*. (Edited by Herbert Feigl and Michael Scriven.) Minneapolis: University of Minnesota Press, 1956. p. 4-37.
39. DAHLSTROM, W. GRANT. "Research in Clinical Psychology: Factor Analytic Contributions." *Journal of Clinical Psychology* 13: 211-20; July 1957.
40. EYSENCK, HANS J. "Criterion-Analysis—An Application of the Hypothetico-Deductive Method in Factor Analysis." *Psychological Review* 57: 38-53; January 1950.

41. FALK, JOHN L. "Issues Distinguishing Idiographic from Nomothetic Approaches to Personality Theory." *Psychological Review* 63: 53-62; January 1956.
42. FEIGL, HERBERT. "Existential Hypotheses." *Philosophy of Science* 17: 35-62; January 1950.
43. FEIGL, HERBERT. "Principles and Problems of Theory Construction in Psychology." *Current Trends in Psychological Theory.* (Edited by Wayne Dennis.) Pittsburgh, Pa.: University of Pittsburgh Press, 1951. p. 179-213.
44. FLANAGAN, JOHN C. "The Use of Comprehensive Rationales in Test Development." *Educational and Psychological Measurement* 11: 151-55; Spring 1951.
45. FLEISHMAN, EDWIN A. "Dimensional Analysis of Psychomotor Abilities." *Journal of Experimental Psychology* 48: 437-54; December 1954.
46. FLEISHMAN, EDWIN A. "Factor Structure in Relation to Task Difficulty in Psychomotor Performance." *Educational and Psychological Measurement* 17: 522-32; Winter 1957.
47. FLEISHMAN, EDWIN A., and HEMPEL, WALTER E., JR. "Factorial Analyses of Complex Psychomotor Performance and Related Skills." *Journal of Applied Psychology* 40: 96-104; April 1956.
48. FRICKE, BENNO G. "A Configural-Content-Intensity Item for Personality Measurement." *Educational and Psychological Measurement* 16: 54-62; Spring 1956.
49. FRUCHTER, BENJAMIN; ROKEACH, MILTON; and NOVAK, EDWIN G. "A Factorial Study of Dogmatism, Opinionation and Related Scales." *Psychological Reports* 4: 19-22; March 1958.
50. FURST, EDWARD J., and FRICKE, BENNO G. "Development and Application of Structured Tests of Personality." *Review of Educational Research* 26: 26-55; February 1956.
51. GAGE, NATHANIEL L. "Logical Versus Empirical Scoring Keys: The Case of the MTAI." *Journal of Educational Psychology* 48: 213-16; April 1957.
52. GETZELS, JACOB W., and WALSH, J. J. *The Method of Paired Direct and Projective Questionnaires in the Study of Attitude Structure and Socialization.* Psychological Monographs, No. 454. Washington, D. C.: American Psychological Association, 1958. 30 p.
53. GINSBERG, ARTHUR. "Hypothetical Constructs and Intervening Variables." *Psychological Review* 61: 119-31; March 1954.
54. GINSBERG, ARTHUR. "Operational Definitions and Theories." *Journal of General Psychology* 52: 223-45; April 1955.
55. GOODENOUGH, FLORENCE L. *Mental Testing.* New York: Reinhart and Co., 1949. 609 p.
56. GUILFORD, JOY P. "Factor Analysis in a Test-Development Program." *Psychological Review* 55: 79-94; March 1948.
57. GUILFORD, JOY P. "New Standards for Test Evaluation." *Educational and Psychological Measurement* 6: 427-38; Winter 1946.
58. GUILFORD, JOY P. *Psychometric Methods.* Second edition. New York: McGraw-Hill Book Co., 1954. p. 341-469.
59. GUILFORD, JOY P. "The Structure of Intellect." *Psychological Bulletin* 53: 267-93; July 1956.
60. GUILFORD, JOY P. "A System of the Psychomotor Abilities." *American Journal of Psychology* 71: 167-74; January 1958.
61. GUILFORD, JOY P.; KETTNER, NORMAN W.; and CHRISTENSEN, PAUL R. "The Nature of the General Reasoning Factor." *Psychological Review* 63: 169-72; May 1956.
62. GUILFORD, JOY P., and ZIMMERMAN, WAYNE S. *Fourteen Dimensions of Temperament.* Psychological Monographs, No. 417. Washington, D. C.: American Psychological Association, 1956. 26 p.
63. GULLIKSEN, HAROLD. "Intrinsic Validity." *American Psychologist* 5: 511-17; October 1950.
64. GULLIKSEN, HAROLD. *Theory of Mental Tests.* New York: John Wiley and Sons, 1950. 486 p.
65. HART, IRWIN. "Maternal Child-Rearing Practices and Authoritarian Ideology." *Journal of Abnormal and Social Psychology* 55: 232-37; September 1957.

66. HEMPEL, CARL G. *Fundamentals of Concept Formation in Empirical Science.* Chicago: University of Chicago Press, 1952. 93 p.
67. HEMPEL, CARL G. "A Logical Appraisal of Operationism." *The Validation of Scientific Theories.* (Edited by Philipp G. Frank.) Boston: Beacon Press, 1957. p. 52-67.
68. HILGARD, ERNEST R. "Intervening Variables, Hypothetical Constructs; Parameters and Constants." *American Journal of Psychology* 71: 238-46; January 1958.
69. HILL, WINFRED F. "Comments on Taylor's 'Drive Theory and Manifest Anxiety.'" *Psychological Bulletin* 54: 490-93; November 1957.
70. HIMELSTEIN, PHILIP. "Goal Setting Rigidity as a Function of Anxiety and Task Ambiguity." *Journal of General Psychology* 58: 69-73; January 1958.
71. JENKINS, JAMES J., and LYKKEN, DAVID T. "Individual Differences." *Annual Review of Psychology.* Vol. 8. Stanford, Calif.: Annual Reviews, 1957. p. 79-112.
72. JENKINS, JOHN G. "Validity for What?" *Journal of Consulting Psychology* 10: 93-98; March 1946.
73. JENSEN, ARTHUR R. "Personality." *Annual Review of Psychology.* Vol. 9. Stanford, Calif.: Annual Reviews, 1958. p. 295-322.
74. JESSOR, RICHARD, and HAMMOND, KENNETH R. "Construct Validity and the Taylor Anxiety Scale." *Psychological Bulletin* 54: 161-70; May 1957.
75. JONES, MARSHALL B. *The Pensacola Z Survey: A Study in the Measurement of Authoritarian Tendency.* Psychological Monographs, No. 452. Washington, D. C.: American Psychological Association, 1957. 19 p.
76. JONES, MARSHALL B. "The Polarity of Psychological Tests." *Journal of Consulting Psychology* 22: 25-29; February 1958.
77. KAESS, WALTER A., and WITRYOL, SAM L. "Positive and Negative Faking on a Forced-Choice Authoritarian Scale." *Journal of Applied Psychology* 41: 333-39; October 1957.
78. KARSON, SAMUEL, and POOL, KENNETH B. "The Construct Validity of the Sixteen Personality Factors Test." *Journal of Clinical Psychology* 13: 245-52; July 1957.
79. KATZELL, RAYMOND A. "Industrial Psychology." *Annual Review of Psychology.* Vol. 8. Stanford, Calif.: Annual Reviews, 1957. p. 237-68.
80. KEEHN, J. D. "Repeated Testing of Four Chronic Schizophrenics on the Bender Gestalt and Wechsler Block Design Tests." *Journal of Clinical Psychology* 13: 179-82; April 1957.
81. KELLY, E. LOWELL. "Theory and Techniques of Assessment." *Annual Review of Psychology.* Vol. 5. Stanford, Calif.: Annual Reviews, 1954. p. 281-310.
82. KELLY, GEORGE A. "The Theory and Technique of Assessment." *Annual Review of Psychology.* Vol. 9. Stanford, Calif.: Annual Reviews, 1958. p. 323-52.
83. KENNY, DOUGLAS T., and GINSBERG, ROSE. "The Specificity of Intolerance and Ambiguity Measures." *Journal of Abnormal and Social Psychology* 56: 300-304; May 1958.
84. KETTNER, NORMAN W.; GUILFORD, JOY P.; and CHRISTENSEN, PAUL R. "A Factor-Analytic Investigation of the Factor Called General Reasoning." *Educational and Psychological Measurement* 16: 438-53; Winter 1956.
85. LAWSON, REED, and MARX, MELVIN H. "Frustration: Theory and Experiment." *Genetic Psychology Monographs* 57: 393-464; May 1958.
86. LEBLANC, MARIA. "Acculturation of Attitude and Personality Among Katangese Women." *Journal of Social Psychology* 47: 257-64; May 1958.
87. LIVERANT, SHEPHARD. *The Use of Rotter's Social Learning Theory in Developing a Personality Inventory.* Psychological Monographs, No. 455. Washington, D. C.: American Psychological Association, 1958. 23. p.
88. LOEVINGER, JANE. "Appendix: Some Substantive Considerations in Construct Validation." Progress Report No. 2, Research Grant M-1213, National Institute of Mental Health. Paper presented to the American Psychological Association, September 1958. St. Louis, Mo.: the author (7209 Pershing Avenue), 1958. 9 p. (Mimeo.)
89. LOEVINGER, JANE. "Objective Tests as Instruments of Psychological Theory." *Psychological Reports* 3: 635-94; Monograph Supplement 9, December 1957.

90. LOEVINGER, JANE. "Some Principles of Personality Measurement." *Educational and Psychological Measurement* 15: 3-17; Spring 1955.
91. MCCLELLAND, DAVID C. "Personality." *Annual Review of Psychology*. Vol. 7. Stanford, Calif.: Annual Reviews, 1956. p. 39-62.
92. MCCLELLAND, DAVID C., and OTHERS. *The Achievement Motive*. New York: Appleton-Century-Crofts, 1953. 384 p.
93. MACCORQUODALE, KENNETH, and MEEHL, PAUL E. "On a Distinction Between Hypothetical Constructs and Intervening Variables." *Psychological Review* 55: 95-107; March 1948.
94. MARX, MELVIN H. "The General Nature of Theory Construction." *Psychological Theory*. (Edited by Melvin H. Marx.) New York: Macmillan Co., 1951. p. 4-19.
95. MARX, MELVIN H. "Hypothesis and Construct." *Psychological Theory*. (Edited by Melvin H. Marx.) New York: Macmillan Co., 1951. p. 112-29.
96. MARX, MELVIN H. "Intervening Variables or Hypothetical Constructs?" *Psychological Review* 58: 235-47; July 1951.
97. MARX, MELVIN H. "Some Suggestions for the Conceptual and Theoretical Analyses of Complex Intervening Variables in Problem-Solving Behavior." *Journal of General Psychology* 58: 115-28; January 1958.
98. MEISSNER, W. W. "Nonconstructural Aspects of Psychological Constructs." *Psychological Review* 65: 143-50; May 1958.
99. MICHAEL, WILLIAM B. "Development of Statistical Methods Especially Useful in Test Construction and Evaluation." *Review of Educational Research* 26: 89-109; February 1956.
100. MICHAEL, WILLIAM B. "Differential Testing of High-Level Personnel." *Educational and Psychological Measurement* 17: 475-90; Winter 1957.
101. MICHAEL, WILLIAM B. "A Suggested Research Approach to the Identification of Psychological Processes Associated with Spatial Visualization Factors." *Educational and Psychological Measurement* 14: 401-406; Summer 1954.
102. MICHAEL, WILLIAM B.; KAISER, HENRY F.; and CLARK, CHERRY ANN. "Research Tools: Statistical Methods." *Review of Educational Research* 27: 498-527; December 1957.
103. MICHAEL, WILLIAM B., and OTHERS. "The Description of Spatial-Visualization Abilities." *Educational and Psychological Measurement* 17: 185-99; Summer 1957.
104. MOSIER, CHARLES I. "A Critical Examination of the Concepts of Face Validity." *Educational and Psychological Measurement* 7: 191-205; Summer 1947.
105. MOSIER, RICHARD D. "Philosophy of the Behavioral Sciences." *Review of Educational Research* 25: 13-24; February 1955.
106. NADELMAN, LORRAINE. "Influence of Concreteness and Accessibility on Concept Thinking." *Psychological Reports* 3: 189-212; June 1957.
107. PEAK, HELEN. "Problems of Objective Observation." *Research Methods in the Behavioral Sciences*. (Edited by Leon Festinger and Daniel Ketz.) New York: Dryden Press, 1953. p. 243-99.
108. REICHENBACH, HANS. "Nomological Statements and Admissible Operations." *Studies in Logic and the Foundations of Mathematics*. Amsterdam: North-Holland Publishing Co., 1954. 140 p.
109. ROBERTS, ALAN H., and JESSOR, RICHARD. "Authoritarianism, Punitiveness and Perceived Social Status." *Journal of Abnormal and Social Psychology* 56: 311-14; May 1958.
110. ROKEACH, MILTON. "The Nature and Meaning of Dogmatism." *Psychological Review* 61: 194-204; May 1954.
111. ROSENBERG, B. G., and ZIMET, CARL N. "Authoritarianism and Aesthetic Choice." *Journal of Social Psychology* 46: 293-97; November 1957.
112. ROSENBLITH, JUDY F. "How Much Invariance Is There in the Relations of 'Prejudice Scores' to Experimental and Attitudinal Variables?" *Psychological Reports* 3: 217-41; June 1957.
113. RYANS, DAVID G. "Notes on the Criterion Problem in Research, with Special Reference to the Study of Teacher Characteristics." *Journal of Genetic Psychology* 91: 33-61; September 1957.
114. SARASON, IRWIN G. "Test Anxiety, General Anxiety and Intellectual Performance." *Journal of Consulting Psychology* 21: 485-90; December 1957.

115. SCHRODER, HAROLD M., and HUNT, DAVID E. *Failure Avoidance in Situational Interpretation and Problem Solving.* Psychological Monographs, No. 432. Washington, D. C.: American Psychological Association, 1957. 22 p.
116. SELLARS, WILFRID. "Concepts as Involving Laws and Inconceivable Without Them." *Philosophy of Science* 15: 287-315; October 1948.
117. SIEGMAN, ARON W. "Cognitive, Affective and Psychopathological Correlates of the Taylor Manifest Anxiety Scale." *Journal of Consulting Psychology* 20: 137-41; April 1956.
118. SILVERMAN, ROBERT E. "The Edwards Personal Preference Schedule and Social Desirability." *Journal of Consulting Psychology* 21: 402-404; October 1957.
119. SPIKER, CHARLES C., and MCCANDLESS, BOYD R. "The Concept of Intelligence and the Philosophy of Science." *Psychological Review* 61: 255-66; July 1954.
120. TAFT, RONALD. "Is the Tolerant Personality Type the Opposite of the Intolerant?" *Journal of Social Psychology* 47: 397-405; May 1958.
121. TAMKIN, ARTHUR S. "An Evaluation of the Construct Validity of Barron's Ego-Strength Scale." *Journal of Clinical Psychology* 13: 156-58; April 1957.
122. TAMKIN, ARTHUR S., and KLETT, C. JAMES. "Barron's Ego Strength Scale: A Replication of an Evaluation of Its Construct Validity." *Journal of Consulting Psychology* 21: 412; October 1957.
123. TITUS, H. EDWIN, and HOLLANDER, EDWIN P. "The California F Scale in Psychological Research: 1950-1955." *Psychological Bulletin* 54: 47-64; January 1957.
124. TRAVERS, ROBERT M. W. "Individual Differences." *Annual Review of Psychology.* Vol. 6. Stanford, Calif.: Annual Reviews, 1955. p. 137-60.
125. TRAVERS, ROBERT M. W. "Rational Hypotheses in the Construction of Tests." *Educational and Psychological Measurement* 11: 128-37; Spring 1951.
126. WALLON, EDWARD J., and WEBB, WILSE B. "The Effect of Varying Degrees of Projection on Test Scores." *Journal of Consulting Psychology* 21: 465-72; December 1957.
127. WEBSTER, HAROLD. "Correcting Personality Scales for Response Sets or Suppression Effects." *Psychological Bulletin* 55: 62-64; January 1958.
128. WITTENBORN, J. RICHARD. "The Theory and Technique of Assessment." *Annual Review of Psychology.* Vol. 8. Stanford, Calif.: Annual Reviews, 1957. p. 331-56.
129. YATES, AUBREY J. "The Validity of Some Psychological Tests of Brain Damage." *Psychological Bulletin* 51: 359-79; July 1954.
130. ZIMMERMAN, WAYNE S. "Hypotheses Concerning the Nature of the Spatial Factors." *Educational and Psychological Measurement* 14: 396-400; Summer 1954.

CHAPTER VIII

Development of Statistical Methods Especially Useful in Test Construction and Evaluation

WILLIAM B. MICHAEL

DURING the period reviewed, a substantially larger number of published papers appeared in statistical methods that are particularly applicable to the analysis, evaluation, and construction of tests than had appeared during the preceding period.

The organization pattern of the current chapter follows essentially that of the corresponding chapter in the February 1956 issue of the REVIEW, "Educational and Psychological Testing" (1), although the order in which problem areas are considered is somewhat different. Moreover, a section concerning the development of statistical models for the analysis of test-taking behavior has been added. As in the corresponding chapter of three years ago, several empirical studies are mentioned that furnish evidence regarding the effectiveness of various statistical procedures when they are applicable to the analysis and evaluation of item and test data. It should be pointed out that developments in factor analysis were reviewed in the chapter on statistical methodology in the December 1957 issue of the REVIEW (2).

Prediction Techniques

During the three-year period under consideration the amount of published research relative to prediction was substantial. Exclusive of pattern and profile analysis which is to be treated as a separate area of research, the attempt to predict a dependent variable from one or more independent variables rested upon the use of both linear and curvilinear models. A great deal of effort was also directed to ascertaining optimal sets of weights for the variables in linear composites when in the absence of a criterion it was desired that each of the variables satisfy certain conditions such as contributing to the variance of the composite in terms of its intended degree of importance.

In a highly meaningful and conceptually oriented article Pickrel (91) wrote in lucid fashion an excellent comprehensive review of the theory and techniques of the classification problem in which he described and evaluated such approaches as the multiple discriminant function, the multiple cutting score, the unique pattern, and multiple correlation and regression. For one who wishes an overview of predictive procedures in the area of multivariate analyses the reviewer knows of no better single source.

Much more specific were the contributions of other writers to multiple regression theory. Through an algebraic development Brogden (8)

showed that criterion estimates predicted from application of usual multiple regression techniques are optimal for classification of personnel, and that for any chosen assignment of men to jobs the sum of multiple regression estimates on the criterion variable is equal to that sum arising from the criterion scores themselves. To maximize the validity of prediction of a test battery under the restriction that the regression weights be non-negative, Lev (57) derived a computational procedure and provided a numerical example. A third paper concerning multiple regression was that of Creager (19) who extended the mathematical procedures for the determination of multiple and partial regression statistics from uniqueness-augmented factor loadings to the general oblique case; he illustrated computational schemes with a numerical example consisting of two factors and seven predictor variables.

To handle the problem of prediction from two independent variables when they are not linearly related or when interaction effects are suspected to exist, Maxwell (80) proposed a regression equation of the second degree, furnished a geometric interpretation by reducing the equation to standard, or canonical, form, and presented a numerical example based on use of orthogonal components in a simple factorial design. Likewise interested in the predictive possibilities of a nonlinear regression model, Saunders (95) expanded upon the contents of an earlier paper (94) concerning the use of the moderator variable that was described in some detail in a previous issue of the REVIEW (1). In his second paper Saunders presented a mathematical basis of moderated regression, cited several examples, and reported a cross-validation analysis in which the differences between correlations based on linear and moderated regression were relatively slight. In a comprehensive treatment of suppressor variables that exert an influence not too unlike that of the moderator variable, Lubin (66) described the rationale of their function in increasing the validity of a linear composite of predictors; he presented several formulas that serve to assist the research worker in deciding whether to employ a suppressor variable.

An unusually important contribution of both theoretical and practical significance to educational measurement was the development by Lord (61) of a regression equation for estimating the true gain realized by an examinee between his initial and final scores on two equivalent forms of a test since the amount of the difference between two observed scores can be misleading in view of the lack of perfect reliability in both. In addition to furnishing a formula for the reliability of the predicted values obtained, Lord presented a numerical example to illustrate the inadequacies involved in the mere subtraction of initial from final scores and included a graphic procedure by which all tested individuals could be grouped relative to the size of their estimated true increments in scores. To demonstrate the usefulness of Lord's equations, Caffrey (11) applied them to the estimation of true growth in the instance of reading scores.

Questioning the plausibility of Lord's restrictive assumption of equality of error variance in initial and final test scores, McNemar (75) derived in a simple and direct manner a regression equation for the prediction of true gains when error variances are not assumed to be equal in the two sets of scores. In addition to this multiple regression equation which makes use of more familiar and easy-to-follow notation than that found in Lord's derivation, McNemar suggested a very simple regressed score method and through use of realistic numerical examples showed that it yields satisfactory approximations to his multiple regression approach, especially in the situation of short-range growth periods. Finally McNemar questioned Lord's statement regarding the additivity of predicted gains in multiple testings and showed that such a circumstance would be atypical.

In another paper closely related to the work of Lord and of McNemar, Garside (34) considered both a linear and curvilinear model in the estimation of the regression of true gains upon initial scores and compared his approach with two other methods.

Employing a procedure intended to overcome several of the limitations involved in obtaining crude gain measures in proficiency, Manning and DuBois (79) carried out with 213 Navy trainees an empirical investigation of the extent to which the same weighted combinations of predictor variables were correlated with crude gain scores, residual gain scores (defined in standard scores as the difference between actual final proficiency z_2 and the final proficiency predicted from initial scores $r_{12}z_1$), and final status scores. The crude gain scores yielded multiple correlations between .14 and .17; the residual gain scores, coefficients between .28 and .40; and the final status scores, indexes between .35 and .52. From logical considerations, the writers concluded that in the training situation the residual gain may be, in many instances, the most meaningful measure of proficiency to be correlated with aptitude test scores.

In the first of three related papers in which test length and testing time were used synonymously, Horst (42) presented and illustrated a method for determining the optimal distribution of testing time among several predictors needed to achieve a maximum index of efficiency in the differential prediction of several criterion variables. Subsequently Horst and MacEwan (44) developed an analogous procedure for multiple absolute prediction and gave a numerical example. To overcome the limitation in the mathematical rationale of two previous papers that the altered time allotment could not approach zero, Horst and MacEwan (45) extended their development in the instance of both multiple differential prediction and multiple absolute prediction to permit the altered time allowance in one or more tests to approach zero. In all three papers iterative procedures are employed to determine, relative to an index of predictive efficiency, the optimal distribution of a newly specified over-all testing time for all predictors when the following are known: the original amount of testing time for each predictor in the battery, the intercor-

relations between potential predictors, their correlations with each of the criterion variables, and their reliabilities.

There were four other papers regarding prediction techniques that were primarily concerned with problems of validity and selection. Cureton (21) dealt with the case when the proportion of individuals in the key category of a dichotomous criterion is a value such as .10 or .90. In the case of curvilinear regression existing for two variables, Perry (89) devised a cutting point theory involving two critical scores on the independent variable such that an individual would place in one of the categories of a dichotomous normally distributed criterion variable at a specified probability level. In order to effect a computational simplification, McCollum and Savard (69) illustrated a direct empirical method for ascertaining the effectiveness of tests in selection that yielded results in relatively close agreement with those furnished by the Taylor-Russell approach. In making an empirical comparison of two methods of test selection and weighting, Lawshe and Patinka (56) demonstrated that a short-cut method for multiple correlation proposed by Jenkins gave results close to the Wherry-Doolittle solution.

Of interest to the research worker in test development and evaluation are several contributions involving weighted composites without a dependent variable. Upon the assumption that the components of a test Y of increased length are parallel forms of test X of unit length, Hoffman (41) derived and compared empirically two formulas that relate the length of a test Y to the weight intended for it relative to test X in the determination of a composite score. In other words, a solution was furnished for ascertaining what the length of a second test should be in relation to another one in a composite in order that their weights would be of predetermined magnitude. Employing two weighting schemes, one in which the ratio of the standard deviation of the augmented test Y to that of the test X of unit length is taken and a second in which the ratio of the standard deviation of the true score of test Y to that of test X is chosen, Hoffman succeeded in the first instance in expressing the weights as a function of both the ratio k of the length of test Y to that of test X and the reliability of test X and also furnished a useful computing diagram. In the second approach it was shown that the weight is equal to k itself. In addition, Hoffman derived a formula from which test reliability could be estimated from knowledge of the ratio k and the standard deviations of the tests of unit and of altered length.

After pointing out inadequacies in arriving at a composite score from either raw scores or standard scores of several measures of the same attribute, Dunnette and Hoggatt (25) outlined an approach to achieve precise weightings according to the importance that one wishes to place on each variable of the composite. Prior to describing both the complex mathematical derivation and an iterative procedure for solution of the system of quadratic equations, the writers demonstrated empirically that

their model assures the desired percent contribution of each variable (rater) to the composite score variance.

Specifying in advance the true score of a composite and the linear independence of the true scores on the different tests within a battery, Woodbury and Lord (111) derived formulas for weighting tests in a battery and for allocating time of test administration such that the reliability of the composite will be a maximum. They demonstrated that except for sign the optimum scoring weight for each test is simply the reciprocal of the standard error of measurement when the test is of unit length.

Among other studies concerned with weighted composites are those of Jones (49) who offered certain refinements in the work of Dingman and Guilford (23) who considered the problem of forming a weighted composite of ratings when a single common factor describes the intercorrelations of raters. In the weighting of personnel data for optimal combination Lawshe and Harris (55) described and illustrated a method of reciprocal averages. From the results of an empirical study concerning the determination of optimal weights for test variables in a composite, Trites and Sells (102) concluded that a unit weighting procedure yields, for practical purposes, essentially the same order of scores as the more cumbersome fractional weighting system. Likewise, from his study of three large samples, Jurgensen (51) concluded that in view of the high degree of correlation between statistically determined and arbitrary weights in employee rating scales, no practical difference would occur in the reliability of estimates if the simpler system should be followed. Criticizing these two articles, McCornack (70) demonstrated mathematically that although two sets of item or test weights in a composite may be highly correlated (even to an extent in excess of .99), the validities of the two keys may differ both to a statistically significant and practically important degree.

Estimation of Test Reliability

As in the past, a substantial amount of diversified research of a theoretical nature appeared in the estimation of test reliability. Clearly no diminution in interest or effort occurred in the study of internal-consistency approaches to reliability. Not to be overlooked was the tendency to relate reliability theory to the models furnished by factor analysis and analysis of variance.

A truly singular contribution to the theory and interpretation of reliability was the penetrating and definitive paper by Tryon (103) who not only examined critically the prevailing assumptions about the nature of measures of individual differences represented by the Spearman-Yule theory of true and error factors and the Brown-Kelley theory of statistically equivalent test samples, but also developed and illustrated numerically his own objective principles of domain sampling which (without

invoking several unnecessarily restrictive assumptions) served as the basis for the derivation of four alternative computing formulas yielding the same numerical estimates of reliability. In his development of operational procedures for the estimation of the reliability of an observed set of X_t scores, Tryon devised a second comparable composite of scores X_t' as a theoretical construct such that n-test samples drawn from this X_t' composite vary on the average as much with respect to the magnitudes of the standard deviations and intercorrelations as would the n-test samples taken from the experimentally available X_t composite. In the instance of both unstratified and stratified composites and domains, the number of test samples, the mean variance, and mean covariances in the observed composite are taken to be equal to those corresponding quantities in the construct composite. In addition the mean of cross covariances between the test samples of X_t and those of X_t' are required to maintain certain relationships to each other relative to the structure of X_t, and in the instance of the unstratified composite the mean of the cross covariances is taken equal to the mean (observed) covariance of test samples from X_t. The familiar index of reliability was shown to be the behavior domain validity of X_t since it represents the degree of correlation between the observed sample and a perfect criterion—the theoretical construct X_t' of infinite length. Finally Tryon concluded that factor postulates underlying much reliability theory constitute a form of orthodoxy that is both unnecessary and superfluous to the understanding of reliability.

After furnishing a succinct and penetrating overview of several published papers related to the Kuder-Richardson reliability formulas, Lord (60) not only presented a new and extremely useful derivation of the Kuder-Richardson formula 21 based on the relations between randomly parallel tests in which the assumption of equal difficulty of items was not necessary, but also developed a formula for the determination of a least upper bound of the reliability coefficient for parallel test forms composed of matched samples of test items. Relative to the formula for the standard error of measurement that served as the basis for his derivation of an estimate of reliability from randomly parallel tests, Lord (58) posed a number of critical questions regarding conditions under which the formula might be used; he demonstrated that in most practical situations parallel tests of the same length might be considered to have, for a given individual, comparable standard errors of measurement. In Lord's formula the amount of standard error of measurement (which is defined as the standard deviation of a single examinee's scores on a large number of parallel forms) is dependent upon only the true score of the examinee (which can be estimated from the obtained score) and the number of items in the test when they are scored 1 or 0. In a third important paper Lord (64) developed a likelihood-ratio significance test for the hypothesis that subsequent to correction for attenuation two

variables measure the same ability or trait, or equivalently that the correlation of the two variables is unity.

The estimation of test reliability through use of analysis-of-variance models was the basis of four important papers. To estimate both a coefficient of internal consistency (an intra-class correlation of responses to the items of an examination reflecting the extent of variance between individuals on one administration) and a coefficient of external consistency (an intra-class correlation representing the stability of responses to the items or of scores received by examinees on different administrations), Moonan (85) developed an experimental model which, when treated by analysis of variance, furnishes a means of obtaining both point and interval estimates of the indexes of external and internal consistency. Subsequently Moonan (86) presented a detailed computational illustration of the method based on the analysis of real data. In ascertaining the effect of a change in length of an examination upon the index of internal consistency Moonan (87) made adaptations in his analysis-of-variance model.

In a comprehensive and detailed paper Burt (10) described and illustrated various ways in which test reliability could be estimated through different modifications of the analysis-of-variance model. Immediately following Burt's article was one in the same journal by Mahmoud (78) who treated the reliability problem primarily in terms of factor theory; related his findings to those of Burt; and concluded that for the illustrative data considered, the factorial approach—especially the group factor model—furnished results superior to those afforded by the analysis of variance.

Questioning earlier work by Cronbach regarding the relationship between the factorial properties of test items for the two situations in which there is or is not a continuous distribution associated with dichotomous item scores, Cotton, Campbell, and Malone (18) concluded for the first case that the proportion of common-factor variance H^2 in a test, which may be described as a function of the intercorrelations among items, is somewhat greater than the well-known coefficient alpha except when only one common factor is present for the items and when the loading of each item in the single factor is inversely proportional to its standard deviation. For the second case, the writers refuted both the existence of a factorial structure of item scores (the product-moment correlations of which constitute a matrix of phi coefficients) and consequently the interpretation of a Kuder-Richardson coefficient $r_{K\text{-}R}$ in terms of factorial properties. Although denying the factorial interpretability of dichotomized scores as basic data, they demonstrated that the magnitude of $r_{K\text{-}R}$ is equal to the value of a coefficient of equivalent H^2phi (a hypothetical test-retest correlation such that specific factor contributions of the underlying distribution are excluded from the self-correlation of items) when the mean variance of items associated with common factors is equal to the average

covariance between items. An empirical study consisting of synthetic test data was included.

For factor scales that are to be lengthened with items that contribute additional variance to the basic unitary factor and additional specific factor variance different from that of any other items or subtests within a scale, Cattell (13) developed formulas for estimating the augmented reliabilities and (internal) validities, and presented a table showing for various multiples of increase in scale length the relationships of reliabilities and validities of the altered tests for two typical levels of .30 and .50 in factor loadings of items (item validities).

In an important empirical study Towner (101) investigated the amount of distortion arising in the application of six methods of estimation of reliability appropriate to a single administration of a test when many of the assumptions underlying use of the methods were not met. Employing four samples of 400 medical students from the freshman through the senior year from whom data were available upon the same achievement test of cancer knowledge, he concluded in light of the marked degree of similarity of the size of the reliability coefficients within each sample that in practice the Kuder-Richardson formula 21 could be used satisfactorily to obtain a quick estimate of test reliability.

Additional contributions were those of Edgerton (26) who developed and illustrated a procedure for estimating the reliability of the average of rankings assigned to individuals; of Cartwright (12) who furnished a rapid nonparametric method for estimation of the reliability of the ratings of several judges; and of Zaccaria, Schmid, and Klubeck (112) who described a simple procedure for the development of equivalent forms of interest or personality inventories.

Evaluation of Sampling Error in Item and Test Analysis

Rather closely related to the theory of estimation of reliability is the stability of item and test statistics associated with the sampling of both individuals and items, respectively, from populations of examinees and of items. Although during the three-year period between August 1, 1955, and July 31, 1958, there was no single contribution comparable in fundamental importance or in scope to the pioneer article by Lord (63) in March 1955 concerning sampling fluctuations arising from the sampling of test items as well as individuals, several significant studies did appear.

Perhaps of greatest interest was the derivation by Keats (54) of a simple formula for the determination of the amount of error variance at a given score level in a test of equivalent items. Representing a small sample estimate of error variance corresponding to the one that Lord (63) devised in the instance of a large sample of test items, Keats's formula, which is independent of fluctuations in the reliability coefficient from population to population for the same test, shows that at a specified score level the amount of error variance stays constant. After introducing

approximate procedures when items are not equivalent, Keats also furnished empirical evidence indicative of the satisfactory results attained from application of these procedures.

Relative to the sampling of items two other papers were noteworthy. For the reliability coefficient based on the use of random halves when all possible splits are sampled for a group of examinees, Lord (62) derived a formula for the sampling variances and showed that when the number of items is large, the estimate of sampling variance furnished exceeds that of the Kuder-Richardson reliability coefficient (formula 20) by a multiplicative factor equal to the number of test items. Although the sampling error of a reliability coefficient based on the use of matched halves of a test would be smaller than the one associated with random halves, it would still be expected to exceed that of the Kuder-Richardson coefficient. In the second paper, which is concerned with probabilities of overlap in item sampling, Anderson and Nuthmann (3) proposed an exact significance test to ascertain whether a sample of m objects, of which k are marked in a designated manner, could be expected to have arisen from a population of N objects, of which a number n bear the same marking as the k objects. Thus in their illustrative example involving a population of 550 *MMPI* items (N objects) from which 72 items (k objects) are chosen on rational grounds, the writers determined that the null hypothesis was supported even when there was only one of the rationally chosen items ($k = 1$) in an empirically obtained sample of 25 items (m objects).

Relative to the sampling of individuals rather than items, Brogden (7) derived a computationally feasible expression for estimating the expected variance, but not the distribution of sampling errors, found in a set of item-criterion correlations (point biserial coefficients) when the items are relatively homogeneous in difficulty. Concerned with the sampling error of individual point biserial coefficients, Perry and Michael (90), in answer to certain valid criticisms about the contents of two of their earlier papers, offered in the instance of large samples new approximations to the determination of the confidence intervals for the coefficient. In three related papers based on a sampling of individuals rather than items McHugh (71, 72, 73) described and illustrated a method of determination of sample size in validation research, proposed an improved formula for estimation of the confidence interval of a true score, and furnished a predictive confidence interval for a validity coefficient.

In an empirical study concerned with the sampling of both items and individuals, Johnson and Lord (48) compared the relative effectiveness of administering either the same items or random samples of different items in the estimation of the mean of a single group of examinees and in the ranking of means of a number of different randomly chosen groups of examinees. They concluded that the unusual procedure of assigning different items to different students not only yielded relatively more con-

sistent mean values than the conventional approach of giving the same items to each examinee, but also furnished a means of saving time in large-scale testing surveys.

Item Selection and Item Analysis Procedures

A great deal of interest was apparent in the development of new procedures in item selection and in item analysis as well as in the modification of techniques already in existence. Although in several of the papers to be reviewed, certain arithmetic savings are involved, the emphasis will rest largely upon the methodological importance of the contribution to item and test analysis. During the three-year period under review there seemed to be a tendency toward a high degree of specialization in most of the papers concerned with item analysis and item selection. One important exception was the significant contribution of Ryans (92) who without employing a single statistical formula presented a comprehensive outline of possible research designs to be used for the selection of items and for the validation of items and scoring keys.

As might be expected, the efforts of several investigators were directed toward the development or modification of indexes of item discrimination. Elaborating upon the contents of an earlier paper (28) which was described in considerable detail in the February 1956 issue of the REVIEW (1), Findley (29) explained and illustrated the use of his easily understood and readily applied discrimination index *D*. Subsequent to deriving and illustrating a new index of item-criterion relationship *lambda*, based upon the ratio of the point biserial coefficient to the maximum point biserial coefficient of the same sign, Clemans (15) pointed out that *lambda* which ranges in value between zero and unity is superior to the biserial or point biserial coefficient in view of its independence from item difficulty, its consistent upper limit of unity, and its sensitivity to any departure from a perfect relationship. After giving a simplified rederivation of the well-known upper and lower 27-percent rule for item discrimination when a normal distribution of criterion scores is assumed along with a constant error of measurement throughout the range of scores, Cureton (22) not only criticized those assumptions but also showed mathematically for a rectangular distribution of test scores that the use of upper and lower thirds constitutes the optimal choice of item analysis. Since many criterion or total test score distributions tend to be platycurtic, it was recommended that the subgroups should probably consist of the upper and lower 29 or 30 percent of the total sample.

Of considerable importance in setting standards for the selection of test items for maximizing test validity and for increasing test homogeneity are the two contributions by Webster (107, 108). In order to increase test homogeneity through item selection, Webster in his first paper proposed a number of techniques, some of which are independent of and

others dependent upon test length, to maximize the degree of homogeneity as measured by the Kuder-Richardson formulas 20 and 21. In addition to deriving exact selection conditions that necessitate only item count information, he furnished applications of his analytic formulations. In his second paper Webster first derived a procedure that describes an exact condition for discarding k items from a pool of potential items such that the residual test will correlate higher with an external criterion than the initial one; he then (with the imposition of certain restrictions on the items to be included in the initial test) outlined for use in test construction a practical method that possesses several advantages over existing techniques.

As a means both of studying the extent to which the validity of multiple-choice items is dependent upon the relationship among the various alternatives and of furnishing a basis upon which the test constructor can make decisions in selecting, preparing, or modifying such items, Cronbach and Merwin (20) developed and illustrated a rather complex model involving the applications of scaling and factor analytic procedures. The computational difficulties involved require the use of electronic computers.

At least four significant papers appeared regarding the attenuation paradox, which is interpreted to mean essentially that as the reliability of a test (as reflected by the average intercorrelation among homogeneous items of comparable difficulty) increases, the validity of the test in terms of its correlation with the common factor underlying it at first rises and then decreases after a certain point. In the first of a series of highly theoretical papers concerning the use of statistics and probability models in item analysis and classification problems that were prepared by mathematical statisticians for the USAF School of Aviation Medicine, Solomon (98) presented a mathematical formulation of the attenuation paradox and furnished several charts to illustrate the paradoxical relationship of validity to reliability in terms of chosen constant levels of item difficulty, equal item intercorrelations, and different numbers of items. In a subsequent paper in the series Sitgreaves (96) formulated a probability model to study item characteristics for a test of a single ability and developed a rather abstract but elegant statistical interpretation of the attenuation paradox in relation to the model.

After pointing out that the "region of the paradox" is substantially reduced when a curvilinear correlation coefficient instead of the usual product-moment coefficient is used, Lord (65) showed that for values of a precision index usually encountered in practical aspects of test development the attenuation paradox can be ignored, urged that greater attention be directed to the study of desired degrees of discriminating power of tests at various levels of ability, and outlined a procedure for determining the optimum difficulty level of items for a test employed in the selection of a specified proportion of examinees when the average item-test biserial correlation is specified. Expressing dissatisfaction with the normal curve

as a model for the distribution of test scores and urging that rank-order data and point distributions be employed, Humphreys (46) demonstrated that if criterion distributions can assume any shape as permitted in the calculation of phi coefficients for interitem correlations (indicating reliability) and of point biserial coefficients for item validities (from which a correlation of sums formula gives a test validity coefficient), no paradox occurs since its locus arises from the fact that one cannot hold constant both the distribution of item difficulties and the shape of the criterion distribution. In describing a sequence of steps that the technician should follow in test construction, Humphreys not only stressed the importance of high item reliability, the maintenance of a specified (though not necessarily high) level of homogeneity despite the possibility of low item-test correlation, and the attainment of the desired distribution of raw scores through varying only item difficulties, but also expressed his preference for a rectangular distribution of scores for a "general purpose" test.

Two other papers concerning theoretical aspects of item analysis with single ability tests appeared. Continuing her work, Sitgreaves (97) proposed a somewhat different model of a more restricted nature in terms of which she arrived at an index h, defined as one minus the minimum expected squared value of an error of estimate in the observed item scores, for the evaluation of the test. Following up Sitgreaves' two studies, Birnbaum (6) applied the Neyman-Pearson and Wald theories of statistical inference and decision making to problems of efficient test design and proposed a logistic function rather than the customary normal ogive to represent the item characteristic curve.

The development of expressions to relate test parameters to item parameters was a central objective of at least three papers. For both equally weighted and differentially weighted items MacLean (74) presented an extremely useful method of deriving from the matrix of item scores familiar statistics descriptive of the test performance of a group of examinees and gave a numerical example. With the presence of a moderate amount of electronic equipment his ingenious formulation would readily permit the completion of conventional item and test analyses within a very short time interval. In developing a theory of item-analysis based on the scoring of items at three levels of appropriateness of response, Michael and Perry (83) furnished formulas that relate item properties to test parameters of mean, variance, reliability, and validity. In transforming test, or criterion, scores to values of 2, 1, 0, −1, and −2 such that 9, 19, 44, 19, and 9 percent of the cases fall into each of five categories, respectively, Webster (109) developed and illustrated easily applied and relatively efficient formulas that furnish estimates of the item-test point biserial correlation, the covariance of original test scores with given items, the variance of test scores, the validity coefficient of a test with an external criterion, and the reliability of a test based on Kuder-Richardson reliability formula 20.

Pertinent to item-selection and item-analysis procedures were the data of three quite different empirical studies. In support of the feasibility of sequential analysis procedures in item selection was the finding by Tiffin and Hudson (100) that this approach is apparently as effective in the realization of test validity and test reliability as the conventional *D*-value item analysis despite the existence of a marked restriction of range in the talent of the groups studied. In evaluating the effect of scoring procedure and length of key upon the validity and reliability of forced-choice tetrads, Harris, Howell, and Newman (37) concluded that (a) positive and positive plus negative weights in scoring keys yield comparable validities relative to the criterion of worker associates' ratings; (b) changes in length of scoring keys containing approximately the best 25, 20, and 15 tetrads do not substantially influence the validity of the forced-choice evaluations; (c) Spearman-Brown predictions of reliability are fairly accurate although they seem to be somewhat more nearly correct when the scoring key is increased by the inclusion of negative weights than by the addition of only positive alternatives; and (d) estimates of reliability tend to decrease when a reduction in the number of scored alternatives occurs. From the study of items in an interest questionnaire embodying three levels of response (like, indifferent, and dislike) that had been validated upon three different samples and cross-validated upon five new samples, Gadel (33) obtained evidence yielding some support for the hypothesis that items showing a high curvilinearity index in the differences of response percentages as derived from upper and lower criterion subgroups show greater shrinkage in validity upon cross validation than do items reflecting linearity in response percentage differences.

Computational Aids to Item Analysis

One of the striking findings in the review of the literature for the last three-year period relative to that of the preceding three years was the marked reduction in the number of articles concerned with computational aids for item analysis such as abacs, charts, tables, and short-cut formulas. It may well be that in light of the increased availability of automatic computers during the past three or four years less need has been felt for practical aids.

However, a few somewhat unrelated articles did appear. Thus, from the ratio of the cross-products of the frequencies in a fourfold table Jenkins (47) showed through use of two tabled corrections how tetrachoric coefficients of correlation could easily be estimated with a mean discrepancy less than .005 even though the splits vary substantially from the medians. To make provision for the estimation of negative tetrachoric coefficients which Jenkins did not consider, Fishman (30) pointed out adaptations that could be conveniently effected in Jenkins' presentation. Another computational short cut for the estimation of a tetrachoric

coefficient was a nonparametric approximation formula devised by Sakoda (93). For the rapid calculation of partial correlation coefficients Michael and Caffrey (81) described the development and use of a set of tables.

To facilitate the calculation of the *D*-statistic in profile analysis, Cleven and Meador (16) explained at length a punched-card procedure. Tucker, DuBois, and Smith (104) gave a detailed account of how to score item punched cards through use of selector networks.

To ascertain the theoretically expected amount of systematic error arising from use of upper and lower 27-percent criterion groups in the estimation of item difficulties for a total criterion sample, Michael, Jones, and Perry (82) explained and illustrated the use of an abac, and from the results of an empirical study Jones and Michael (50) recommended that corrections in estimates of item difficulty afforded by the abac be considered when item validities of .50 or higher are found. Likewise, in an empirical investigation involving the use of upper and lower 27-percent groups for determination of item difficulty, Clark (14) reported that his procedure became progressively less satisfactory as an increase occurred in item discriminating quality.

Transformation of Scale Values

Seven papers concerned the transformation of scores. Probably the most significant contribution was the maximum likelihood solution proposed by Lord (59) for equating two tests U and V that have been administered to different randomly chosen groups of examinees when a third anchor test W has been given simultaneously to each group. In a comprehensive empirical study concerning the effects of sampling error that arise from equating scales (essentially parallel tests) administered to nonoverlapping groups, Karon (53) compared (under both stratified and random sampling) Lord's solution with three other equating techniques (the conventional mean and sigma method, the equi-percentile procedure, and a standard reference group approach also requiring an anchor test). Karon concluded generally that Lord's formulation was the most satisfactory although its sampling error was insignificantly larger than that of the other anchor-oriented approach which gave slightly biased results. In addition to showing the largest amount of error to be associated with the equi-percentile procedure, Karon also found that sampling error was (a) smaller for methods making use of anchor tests than for those that do not, (b) smaller for equating of scores close to the mean of the total population than for those scores farther removed from the mean, and (c) not diminished under conditions of stratification when the two methods embodying anchor tests were applied. Making use of three assumptions involving linearity and employing total score as the criterion in item analysis, Swineford and Fan (99) described a method for converting scores on one form of a test to those on another (parallel)

form through use of item statistics and suggested that their proposed "item method of conversion" would be applicable to the situation in which a subset of items is common to two test forms that had been administered to two different groups. Concerned with the problem of trying to equate scores on two nonparallel tests measuring different functions, such as the familiar *ACE Psychological Examination* and the *College Board Scholastic Aptitude Test,* Angoff (4) pointed out that sources of error arise from methodological distinctions in definitions of comparability, from the existence of ability differences in the normative populations, and from the presence of differential selection effects. He then proceeded to demonstrate empirically that the more nearly homogeneous the groups compared and the more nearly similar the functions measured in the two tests, the less will be the extent of discrepancies between corresponding sets of converted scores.

In making a critical examination of the application of Thurstone's method of absolute scaling to problems of item scaling and of score scaling, Fan (27) concluded from use of both empirical and fictitious data that the fundamental assumption of the identity between the two equations for test score conversion and item difficulty conversion that necessarily implies the equivalence of slopes and intercepts of the corresponding lines representing the two equations is false unless the two groups of examinees being compared are similar in the characteristic measured. From a somewhat more substantive point of interest, Whiteman and Jastak (110) applied Thurstone's methods of absolute scaling to three subtests of the *Wechsler-Bellevue Scale* over the age range from 10 through 64 in an attempt to overcome sampling biases and inequality of units in measures.

The remaining paper concerning transformation of scores contained a development by Kaiser (52) of a modified stanine score in which the standard deviation is 2.00 instead of the more familiar value of 1.96.

Profile and Pattern Analysis

In the increasingly specialized area of prediction known as profile and pattern analysis several noteworthy contributions appeared, largely at a theoretical level. Constituting an important analytic achievement, if not a major breakthrough, was the highly readable paper by Lubin and Osburn (67) who proposed a comprehensive theory of pattern analysis for the prediction of a normally distributed criterion from dichotomous items along with the needed significance tests. In addition to furnishing a technique for computation of the configural scale as a polynomial function that in the least-squares sense was shown to possess maximum validity, Lubin and Osburn described *F*-ratio significance tests to ascertain whether the validity of the configural scale is greater than zero, whether its validity is significantly larger than that of the total score

derived from unweighted items, whether the relationship between item scores and the quantitative criterion is linear or nonlinear, whether the addition of certain items will contribute to a gain in the validity of the configural scale, and whether nonlinear terms are necessary in addition to linear ones to yield maximum validity. In a rather closely related paper Osburn and Lubin (88) showed how the configural scale could be used to furnish an exact statistical test as to whether a test-scoring technique such as the multiple-regression, multiple cut-off, or total (unweighted) score approach would yield optimal validity. Because of the almost prohibitively large number of parameters involved in the determination of regression coefficients associated with consideration of all possible answer patterns the method would be appropriate, as the writers mentioned, only in those situations in which the number of items must be extremely small compared with the number of examinees.

In a comprehensive article of more than 40 pages McQuitty (77) reviewed and related to theories of organization of psychological test behavior three kinds of pattern-analytic methods appropriate to unordered data and then applied to Air Force test and criterion measures a dual-pattern method in which patterns of scores on different criteria are determined in conjunction with corresponding patterns of scores on tests, or predictors. His results indicated that his new method yields coefficients of dependability of prediction approximately equal to those of linear models and that in pattern analysis complex tests seem to be relatively more effective than simpler ones in predicting criteria. Somewhat earlier McQuitty (76), without limiting the treatment of his configurational data to placement on linear continua, developed and illustrated a general method of pattern analysis referred to as agreement analysis in which persons can be classified in terms of their predominant pattern of response to test items. Subjecting McQuitty's data to factor analysis, Watson (105) cited what he believed to be certain interpretative advantages that his factorial approach affords.

After reviewing succinctly recent work in configural analysis and pointing out certain inadequacies, Fricke (32) proposed for personality measurement a configural-content-intensity item consisting of a pair of statements. After responding true or false to each of the two statements, the examinee indicates for which one he feels the stronger—a format permitting eight scorable response configurations that Fricke believes may increase the sensitivity of measurement of personality dimensions. Likewise pertinent to personality measurements were the analytic demonstrations and numerical application by Horst (43) of a scheme of configural scoring to test items, and the development of a system of configural analysis by Du Mas (24) that in being less wasteful of available test data would potentially permit a marked reduction in the number of items in a scale, as in the instance of finding the 10 or 15 most serviceable items of the 500 or more in the *MMPI*.

Believing that neither a single index of profile similarity nor a measure of geometric similarity can be expected to yield any genuine psychological utility, Lykken (68) proposed the hypothesis of pattern analysis that various psychological criterion variables can be estimated best through employment of nonlinear joint functions of those test variables making up a given profile. To implement his hypothesis, Lykken described and illustrated a procedure based on the use of the multiple-eta statistic that he believes will furnish a prediction of the criterion dimension, a measure of the degree and significance of the predicted values, and an assessment of the amount of similarity and dissimilarity in the profiles relative to the criterion variable.

Two other theoretical articles concerning profile analysis appeared, both of which are related to factor theory. In his examination of two previously proposed measures of profile similarity for comparing two individuals, Harris (36) expressed the Euclidean distance measures of pairs of persons in matrix notation and made a distinction between the two measures in terms of two different definitions of uncorrelated variables that were associated with an inverse transformation and with a principal-axis transformation. Using matrix notation, Gibson (35) pointed out the identity existing both in the geometric problem and in the solutions found in Lazarsfeld's latent structure model and in Cattell's proportional profile approach. Gibson then proceeded to adapt T. W. Anderson's latent structure formulation to proportional profiles as a possible solution to the communality and rotational problems in factor analysis.

From a somewhat more practical point of view Fricke (31) described for the prediction of academic achievement a relatively simple coded profile method in which he considered simultaneously, rather than individually, the magnitudes and patterns of relationships of two or more test scores; from his illustrative data he concluded that his approach affords certain advantages over the traditional multiple-regression techniques. In another empirical investigation involving artificial data consisting of eight measures on three known groups of geometric forms Helmstadter (38) compared about a dozen methods of estimating profile similarity. Examining the proportion of successful classifications in a cross-validation sample, he found results significantly better than chance, with some methods more successful than others.

Although important strides have been made in the development and refinement of statistical models in pattern and profile analysis, the results from empirical studies have been disappointing to the reviewer. It would seem that use of the traditional linear model in multiple regression is likely to serve almost as well as the application of more sophisticated pattern and profile techniques that involve elaborate and time-consuming calculations. Perhaps in the instance of profile analysis in which difference scores or functions of difference scores are frequently obtained, the reliability is so likely to be attenuated that the potentiality for highly predictive validities, especially upon cross validation, is seriously limited.

Analysis of Test-Taking Behavior

Important during the past three years were studies of the influence on item and test properties of test-taking attitude (e.g., response sets) and partial knowledge (reflected by the tendency to guess). Using two different models, Helmstadter (39) derived several formulas from which separate set and content components of a test score can be obtained, compared empirically the two distinct approaches relative to an ability test administered to 62 graduate students in journalism, and suggested the importance of using an external criterion along with other evidence in the selection of the model to be used. In attempting to correct personality scales for response sets or suppression effects through statistical means, Webster (106) developed an equation to estimate a score for each examinee independent of his response-set score.

To obtain a measure of the gambling response-set associated with application of the correction formula for guessing of which the examinees are apprized, Ziller (113) developed a formula of risk acceptance representing theoretically a ratio of the number of items upon which the examinee guesses to the total number of items he does not know, but upon which he could guess. It is apparent that the risk index is functionally related to the number of items marked incorrectly and to the number of items omitted and hence that the portion of test variance attributable to risk acceptance is dependent upon the difficulty level of the items. Also interested in the problem of correction for guessing and in other factors in the test situation influencing the responses of examinees to items, Brownless and Keats (9) proposed a method involving 10 different types of item response categories relative to two test administrations and gave a numerical example to illustrate application of some of their formulas for describing the frequencies of occurrence of different combinations of responses.

In their detailed proposal of an experimental response method for multiple-choice items based on the theory of partial information in which the examinee understands that he will receive one point of credit for every wrong alternative (distractor) he crosses out and $1\text{-}k$ points if he deletes the correct answer where k is the number of alternatives to the item, Coombs, Milholland, and Womer (17) furnished a score scale ranging from $1\text{-}k$ to $k\text{-}1$ points for each item and carried out an empirical study with three different types of power tests in which they compared the experimental method with the conventional formula for correction of chance successes. In addition to obtaining clear-cut evidence of the presence of partial information in the selection of responses to multiple-choice items, they found that for the tests used an average increment occurred in the reliability to the extent equivalent to about a 20-percent increase in the effective length of a test and that at least as many discriminations between individuals in the test score distribution could be effected with the new approach as with the conventional scoring formula.

To examine the reliability of additional test discriminations effected by having examinees respond in specified ways to each alternative in multiple-choice items (e.g., by having examinees judge the relative degree of accuracy of each of several alternatives) and by weighting each response as to its appropriateness or inappropriateness according to a prescribed plan, Milholland (84) concluded from the formula he derived that increasing the range of scores from the usual $n + 1$ score categories where n is the number of items scored 0 or 1, to $an + 1$ categories where a is the number of item choices each given a score, would not maintain a constant standard error of measurement unless the reliability of the test was substantially augmented.

Other Contributions

Of the more than 100 papers reviewed only two could not be classified in any one of the previous eight divisions of this chapter. Two specific applications of correlation to problems in test analysis were developed. Hills (40) presented formulas for obtaining estimates of the over-all correlations within several groups on the same two variables; in light of possible differences in the reliability estimates of measures in the various subgroups that might be associated with different score variances he furnished corresponding formulas permitting a correction for attenuation. For the estimation of the nonspurious correlation of a part of a test with the total test, rather than of its correlation with the remaining items on the total test, Angoff (5) provided formulas that allow the subtest to consist of items either nonparallel or parallel in form to the total test.

Bibliography

1. AMERICAN EDUCATIONAL RESEARCH ASSOCIATION. "Educational and Psychological Testing." *Review of Educational Research* 26: 5-109; February 1956.
2. AMERICAN EDUCATIONAL RESEARCH ASSOCIATION. "Methodology of Educational Research." *Review of Educational Research* 27: 421-547; December 1957.
3. ANDERSON, NORMAN H., and NUTHMANN, CONRAD F. "An Exact Method for Calculating Probabilities of Overlap in Item Sampling." *Psychological Reports* 1: 317-18; December 1955.
4. ANGOFF, WILLIAM H. "The 'Equating' of Non-Parallel Tests." *Journal of Experimental Education* 25: 241-47; March 1957.
5. ANGOFF, WILLIAM H. "A Note on the Estimation of Nonspurious Correlations." *Psychometrika* 21: 295-97; September 1956.
6. BIRNBAUM, ALLAN. *Probability and Statistics in Item Analysis and Classification Problems: Efficient Design and Use of Tests of Mental Ability for Various Decision-Making Problems.* U. S. Air Force School of Aviation Medicine Report Number 58-16. Randolph Air Force Base, Texas: Air University, November 1957. 25 p.
7. BROGDEN, HUBERT E. "The Expected Variance of the Sampling Errors for a Set of Item-Criterion Correlations." *Psychometrika* 22: 75-78; March 1957.
8. BROGDEN, HUBERT E. "Least Squares Estimates and Optimal Classification." *Psychometrika* 20: 249-52; September 1955.
9. BROWNLESS, VERA T., and KEATS, JOHN A. "A Retest Method of Studying Partial Knowledge and Other Factors Influencing Item Response." *Psychometrika* 23: 67-73; March 1958.

10. BURT, SIR CYRIL. "Test Reliability Estimated by Analysis of Variance." *British Journal of Statistical Psychology* 8: 103-18; November 1955.
11. CAFFREY, JOHN G. "'Estimated True Growth'—Lord's Equations Applied to Reading Test Data." *California Journal of Educational Research* 7: 178-82; September 1956.
12. CARTWRIGHT, DESMOND S. "A Rapid Non-Parametric Estimate of Multi-Judge Reliability." *Psychometrika* 21: 17-29; March 1956.
13. CATTELL, RAYMOND B. "Formulae and Table for Obtaining Validities and Reliabilities of Extended Factor Scales." *Educational and Psychological Measurement* 17: 491-98; Winter 1957.
14. CLARK, EDWARD L. "Item Difficulties Based on End Segments." *Journal of Educational Psychology* 48: 457-59; November 1957.
15. CLEMANS, WILLIAM V. "An Index of Item-Criterion Relationship." *Educational and Psychological Measurement* 18: 167-72; Spring 1958.
16. CLEVEN, WALTER A., and MEADOR, B. J. "Punched Card Calculation of the D Statistic." *Educational and Psychological Measurement* 17: 142-48; Spring 1957.
17. COOMBS, CLYDE H.; MILHOLLAND, JOHN E.; and WOMER, FRANK B. "The Assessment of Partial Knowledge." *Educational and Psychological Measurement* 16: 13-37; Spring 1956.
18. COTTON, JOHN W.; CAMPBELL, DONALD T.; and MALONE, R. DANIEL. "The Relationship Between Factorial Composition of Test Items and Measures of Test Reliability." *Psychometrika* 22: 347-57; December 1957.
19. CREAGER, JOHN A. "General Resolution of Correlation Matrices into Components and Its Utilization in Multiple and Partial Regression." *Psychometrika* 23: 1-8; March 1958.
20. CRONBACH, LEE J., and MERWIN, JACK C. "A Model for Studying the Validity of Multiple-Choice Items." *Educational and Psychological Measurement* 15: 337-52; Winter 1955.
21. CURETON, EDWARD E. "Recipe for a Cookbook." *Psychological Bulletin* 54: 494-97; November 1957.
22. CURETON, EDWARD E. "The Upper and Lower Twenty-Seven Per Cent Rule." *Psychometrika* 22: 293-96; September 1957.
23. DINGMAN, HARVEY F., and GUILFORD, JOY P. "A New Method for Obtaining Weighted Composite of Ratings." *Journal of Applied Psychology* 38: 305-307; October 1954.
24. DU MAS, FRANK. "Concept of the Intratest and Some Implications for Psychometric Theory." *Psychological Reports* 4: 187-92; June 1958.
25. DUNNETTE, MARVIN D., and HOGGATT, AUSTIN C. "Deriving a Composite Score from Several Measures of the Same Attribute." *Educational and Psychological Measurement* 17: 423-34; Autumn 1957.
26. EDGERTON, HAROLD A. "Estimation of the Reliability of Average of Rankings." *Journal of Applied Psychology* 41: 324; October 1957.
27. FAN, CHUNG-TEH. "On the Applications of the Method of Absolute Scaling." *Psychometrika* 22: 175-83; June 1957.
28. FINDLEY, WARREN G. "A Rationale for Evaluation of Item Discrimination Statistics." (Abstract) *American Psychologist* 9: 365-66; August 1954.
29. FINDLEY, WARREN G. "A Rationale for Evaluation of Item Discrimination Statistics." *Educational and Psychological Measurement* 16: 175-80; Summer 1956.
30. FISHMAN, JOSHUA A. "A Note on Jenkins 'Improved Method for Tetrachoric *r*.'" *Psychometrika* 21: 305; September 1956.
31. FRICKE, BENNO G. "A Coded Profile Method for Predicting Achievement." *Educational and Psychological Measurement* 17: 98-104; Spring 1957.
32. FRICKE, BENNO G. "A Configural-Content-Intensity Item for Personality Measurement." *Educational and Psychological Measurement* 16: 54-62; Spring 1956.
33. GADEL, MARGUERITE S. "The Relationship of Item Validity Shrinkage to Curvilinearity of Response Distributions." *Educational and Psychological Measurement* 18: 145-52; Spring 1958.
34. GARSIDE, R. F. "The Regression of Gains upon Initial Scores." *Psychometrika* 21: 67-77; March 1956.
35. GIBSON, WILFRED A. "Proportional Profiles and Latent Structure." *Psychometrika* 21: 135-44; June 1956.

36. HARRIS, CHESTER W. "Characteristics of Two Measures of Profile Similarity." *Psychometrika* 20: 289-97; December 1955.
37. HARRIS, FRANK J.; HOWELL, MARGARET A.; and NEWMAN, SIDNEY H. "Forced Choice Tetrads-Effect of Scoring Procedure and Key Length on Validity and Reliability." *Educational and Psychological Measurement* 16: 454-64; Winter 1956.
38. HELMSTADTER, GERALD C. "An Empirical Comparison of Methods for Estimating Profile Similarity." *Educational and Psychological Measurement* 17: 71-82; Spring 1957.
39. HELMSTADTER, GERALD C. "Procedures for Obtaining Separate Set and Content Components of a Test Score." *Psychometrika* 22: 381-93; December 1957.
40. HILLS, JOHN R. "Within-Group Correlations and Their Correction for Attenuation." *Psychological Bulletin* 54: 131-33; March 1957.
41. HOFFMAN, PAUL J. "Predetermination of Test Weights." *Psychometrika* 23: 85-92; March 1958.
42. HORST, PAUL. "Optimal Test Length for Maximum Differential Prediction." *Psychometrika* 21: 51-66; March 1956.
43. HORST, PAUL. "The Uniqueness of Configural Test Item Scores." *Journal of Clinical Psychology* 13: 107-14; April 1957.
44. HORST, PAUL, and MACEWAN, CHARLOTTE. "Optimal Test Length for Maximum Absolute Prediction." *Psychometrika* 21: 111-24; June 1956.
45. HORST, PAUL, and MACEWAN, CHARLOTTE. "Optimal Test Length for Multiple Prediction: The General Case." *Psychometrika* 22: 311-24; December 1957.
46. HUMPHREYS, LLOYD G. "The Normal Curve and the Attenuation Paradox in Test Theory." *Psychological Bulletin* 53: 472-76; November 1956.
47. JENKINS, W. L. "An Improved Method for Tetrachoric *r*." *Psychometrika* 20: 253-58; September 1955.
48. JOHNSON, M. CLEMENS, and LORD, FREDERIC M. "An Empirical Study of the Stability of a Group Mean in Relation to the Distribution of Test Items Among Students." *Educational and Psychological Measurement* 18: 325-29; Summer 1958.
49. JONES, MARSHALL B. "Composite Ratings and the Case of Unit Rank." *Journal of Applied Psychology* 41: 198-200; June 1957.
50. JONES, ROBERT A., and MICHAEL, WILLIAM B. "An Empirical Study of Systematic Errors in Estimates of Item Difficulty Obtained from Use of Upper and Lower 27 Per Cent Criterion Groups." *Educational and Psychological Measurement* 17: 131-35; Spring 1957.
51. JURGENSEN, CLIFFORD E. "Item Weights in Employee Rating Scales." *Journal of Applied Psychology* 39: 305-307; October 1955.
52. KAISER, HENRY F. "A Modified Stanine Scale." *Journal of Experimental Education* 26: 261; March 1958.
53. KARON, BERTRAM P. "The Stability of Equated Test Scores." *Journal of Experimental Education* 24: 181-95; March 1956.
54. KEATS, JOHN A. "Estimation of Error Variances of Test Scores." *Psychometrika* 22: 29-41; March 1957.
55. LAWSHE, CHARLES H., and HARRIS, D. H. "The Method of Reciprocal Averages in Weighting Personnel Data." *Educational and Psychological Measurement* 18: 331-36; Summer 1958.
56. LAWSHE, CHARLES H., and PATINKA, PAUL J. "An Empirical Comparison of Two Methods of Test Selection and Weighting." *Journal of Applied Psychology* 42: 210-12; June 1958.
57. LEV, JOSEPH. "Maximizing Test Battery Prediction When the Weights Are Required To Be Non-Negative." *Psychometrika* 21: 245-52; September 1956.
58. LORD, FREDERIC M. "Do Tests of the Same Length Have the Same Standard Errors of Measurement?" *Educational and Psychological Measurement* 17: 510-21; Winter 1957.
59. LORD, FREDERIC M. "Equating Test Scores—A Maximum Likelihood Solution." *Psychometrika* 20: 193-200; September 1955.
60. LORD, FREDERIC M. "Estimating Test Reliability." *Educational and Psychological Measurement* 15: 325-36; Winter 1955.
61. LORD, FREDERIC M. "The Measurement of Growth." *Educational and Psychological Measurement* 16: 421-37; Winter 1956.

62. LORD, FREDERIC M. "Sampling Error Due to Choice of Split in Split-Half Reliability Coefficients." *Journal of Experimental Education* 24: 245-49; March 1956.
63. LORD, FREDERIC M. "Sampling Fluctuations Resulting from the Sampling of Test Items." *Psychometrika* 20: 1-22; March 1955.
64. LORD, FREDERIC M. "A Significance Test for the Hypothesis That Two Variables Measure the Same Trait Except for Errors of Measurement." *Psychometrika* 22: 207-20; September 1957.
65. LORD, FREDERIC M. "Some Perspectives on 'The Attenuation Paradox in Test Theory.'" *Psychological Bulletin* 52: 505-10; November 1955.
66. LUBIN, ARDIE. "Some Formulae for Use with Suppressor Variables." *Educational and Psychological Measurement* 17: 286-96; Summer 1957.
67. LUBIN, ARDIE, and OSBURN, HOBART G. "A Theory of Pattern Analysis for the Prediction of a Quantitative Criterion." *Psychometrika* 22: 63-73; March 1957.
68. LYKKEN, DAVID T. "A Method of Actuarial Pattern Analysis." *Psychological Bulletin* 53: 102-107; January 1956.
69. MCCOLLUM, IVAN N., and SAVARD, DAVID A. "A Simplified Method of Computing the Effectiveness of Tests in Selection." *Journal of Applied Psychology* 41: 243-46; August 1957.
70. MCCORNACK, ROBERT L. "A Criticism of Studies Comparing Item-Weighting Methods." *Journal of Applied Psychology* 40: 343-44; October 1956.
71. MCHUGH, RICHARD B. "Determining Sample Size in Validation Research." *Educational and Psychological Measurement* 17: 136-41; Spring 1957.
72. MCHUGH, RICHARD B. "The Interval Estimation of a True Score." *Psychological Bulletin* 54: 73-74; January 1957.
73. MCHUGH, RICHARD B. "A Predictive Confidence Interval for the Validity Coefficient." *Journal of Experimental Education* 24: 323-24; June 1956.
74. MACLEAN, ANGUS G. "Properties of the Item Score Matrix." *Psychometrika* 23: 47-53; March 1958.
75. MCNEMAR, QUINN. "On Growth Measurement." *Educational and Psychological Measurement* 18: 47-55; Spring 1958.
76. MCQUITTY, LOUIS L. "Agreement Analysis: Classifying Persons by Predominant Patterns of Responses." *British Journal of Statistical Psychology* 9: 5-16; May 1956.
77. MCQUITTY, LOUIS L. "Isolating Predictor Patterns Associated with Major Criterion Patterns." *Educational and Psychological Measurement* 17: 3-42; Spring 1957.
78. MAHMOUD, A. F. "Test Reliability in Terms of Factor Theory." *British Journal of Statistical Psychology* 8: 119-35; November 1955.
79. MANNING, WINTON H., and DUBOIS, PHILIP H. "Gain in Proficiency as a Criterion in Test Validation." *Journal of Applied Psychology* 42: 191-94; June 1958.
80. MAXWELL, A. E. "Contour Analysis." *Educational and Psychological Measurement* 17: 347-60; Autumn 1957.
81. MICHAEL, WILLIAM B., and CAFFREY, JOHN G. "Tables To Facilitate Computation of Partial Correlation Coefficients." *Educational and Psychological Measurement* 16: 232-36; Summer 1956.
82. MICHAEL, WILLIAM B.; JONES, ROBERT A.; and PERRY, NORMAN C. "An ABAC for Estimating Certain Systematic Error." *California Journal of Educational Research* 8: 83-86; March 1957.
83. MICHAEL, WILLIAM B., and PERRY, NORMAN C. "A Theory of Item-Analysis Based on the Scoring of Items at Three Levels of Appropriateness of Response." *Educational and Psychological Measurement* 15: 404-15; Winter 1955.
84. MILHOLLAND, JOHN E. "The Reliability of Test Discriminations." *Educational and Psychological Measurement* 15: 362-70; Winter 1955.
85. MOONAN, WILLIAM J. "An Analysis of Variance Method for Determining the External and Internal Consistency of an Examination." *Journal of Experimental Education* 24: 239-44; March 1956.
86. MOONAN, WILLIAM J. "Computational Illustrations of the Internal and External Consistency Analysis of Examination Responses." *Journal of Experimental Education* 25: 181-90; March 1957.
87. MOONAN, WILLIAM J. "The Effect of Changing the Length of an Examination on the Index of Internal Consistency." *Journal of Experimental Education* 26: 209-15; March 1958.

88. OSBURN, HOBART G., and LUBIN, ARDIE. "The Use of Configural Analysis for the Evaluation of Test Scoring Methods." *Psychometrika* 22: 359-71; December 1957.
89. PERRY, NORMAN C. "A Cutting Point Theory for Curvilinear Regression." *Psychological Reports* 3: 78; March 1957.
90. PERRY, NORMAN C., and MICHAEL, WILLIAM B. "A Note Concerning the Reliability of a Point Biserial Coefficient for Large Samples." *Educational and Psychological Measurement* 18: 139-43; Spring 1958.
91. PICKREL, EVAN W. "Classification Theory and Techniques." *Educational and Psychological Measurement* 18: 37-46; Spring 1958.
92. RYANS, DAVID G. "Research Designs for the Empirical Validation of Tests and Inventories." *Educational and Psychological Measurement* 17: 175-84; Summer 1957.
93. SAKODA, JAMES M. "A Nonparametric Approximation Formula for the Tetrachoric *r*." (Abstract) *American Psychologist* 12: 457; July 1957.
94. SAUNDERS, DAVID R. "The 'Moderator Variable' as a Useful Tool in Prediction." *Proceedings of the 1954 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1955. p. 54-58.
95. SAUNDERS, DAVID R. "Moderator Variables in Prediction." *Educational and Psychological Measurement* 16: 209-22; Summer 1956.
96. SITGREAVES, ROSEDITH. *Probability and Statistics in Item Analysis and Classification Problems: A Statistical Formulation of the Attenuation Paradox in Test Theory.* U. S. Air Force School of Aviation Medicine Report No. 57-1. Randolph Air Force Base, Texas: Air University, August 1956. 18 p.
97. SITGREAVES, ROSEDITH. *Probability and Statistics in Item Analysis and Classification Problems: Optimal Test Design in a Special Training Situation.* U. S. Air Force School of Aviation Medicine Report No. 57-117. Randolph Air Force Base, Texas: Air University, September 1957. 35 p.
98. SOLOMON, HERBERT. *Probability and Statistics in Item Analysis and Classification Problems: Probability and Statistics in Psychometric Research with Special Regard to Item Analysis and Classification Techniques.* U. S. Air Force School of Aviation Medicine Research Bulletin No. 56-88. Randolph Air Force Base, Texas: Air University, 1956. 32 p.
99. SWINEFORD, FRANCES, and FAN, CHUNG-TEH. "A Method of Score Conversion Through Item Statistics." *Psychometrika* 22: 185-88; June 1957.
100. TIFFIN, JOSEPH, and HUDSON, TERRANCE W. "Comparison of Sequential and Conventional Item Analysis When Used with Primary Groups Varying in Size and Composition." *Educational and Psychological Measurement* 16: 333-44; Autumn 1956.
101. TOWNER, LEONARD W. "Reliability Coefficients Obtained Under Varying Degrees of Deviation from Theoretically Perfect Conditions." *Educational and Psychological Measurement* 16: 345-51; Autumn 1956.
102. TRITES, DAVID K., and SELLS, SAUL B. "A Note on Alternative Methods for Estimating Factor Scores." *Journal of Applied Psychology* 39: 455-56; December 1955.
103. TRYON, ROBERT C. "Reliability and Behavior Domain Validity: Reformulation and Historical Critique." *Psychological Bulletin* 54: 229-49; May 1957.
104. TUCKER, LEDYARD R.; DUBOIS, PHILIP H.; and SMITH, THOMAS L., JR. "Scoring Item Punched Cards by Selector Networks." *Educational and Psychological Measurement* 16: 237-43; Summer 1956.
105. WATSON, H. E. "Agreement Analysis—A Note on Professor McQuitty's Article." *British Journal of Statistical Psychology* 9: 17-20; May 1956.
106. WEBSTER, HAROLD. "Correcting Personality Scores for Response Sets or Suppression Effects." *Psychological Bulletin* 55: 62-64; January 1958.
107. WEBSTER, HAROLD. "Item Selection Methods for Increasing Test Homogeneity." *Psychometrika* 22: 395-403; December 1957.
108. WEBSTER, HAROLD. "Maximizing Test Validity by Item Selection." *Psychometrika* 21: 153-64; June 1956.
109. WEBSTER, HAROLD. "Transformed Statistics for Use in Test Construction." *Psychological Bulletin* 53: 488-92; November 1956.
110. WHITEMAN, MARTIN, and JASTAK, JOSEPH. "Absolute Scaling of Tests for Different Age Groupings of a State-Wide Sample." *Educational and Psychological Measurement* 17: 338-46; Autumn 1957.

111. WOODBURY, MAX A., and LORD, FREDERIC M. "The Most Reliable Composite with a Specified True Score." *British Journal of Statistical Psychology* 9: 21-28; May 1956.
112. ZACCARIA, MICHAEL A.; SCHMID, JOHN, JR.; and KLUBECK, S. "A Simple Procedure for Developing Equivalent Forms of Interest or Personality Questionnaires." *Psychological Reports* 1: 37-41; March 1955.
113. ZILLER, ROBERT C. "A Measure of the Gambling Response-Set in Objective Tests." *Psychometrika* 22: 289-92; September 1957.

Index to Volume XXIX, No. 1

Page citations are made to single pages; these are often the beginning of a chapter, section, or running discussion dealing with the topic.

REVIEW OF EDUCATIONAL RESEARCH

The REVIEW has been published five times a year, beginning in 1931. The issues that have appeared to date are classified under the following headings. Some of these headings are rather broad; thus ADMINISTRATION includes issues devoted to research on legal and fiscal aspects of education. SPECIAL PROGRAMS includes issues devoted to health, vocational education, and the like, as well as issues considering various levels of education, such as adult education.

Single copies prior to 1949, $1; 1949 through June 1957, $1.50; October 1957 to date, $2 each. Discounts on quantity orders. Orders should be sent to 1201 Sixteenth Street, N. W., Washington 6, D. C.

ADMINISTRATION: I:3 (June 1931); II:2 (April 1932); II:5 (December 1932); III:5 (December 1933); IV:4 (October 1934); V:2 (April 1935); V:4 (October 1935); VII:4 (October 1937); VIII:2 (April 1938); VIII:4 (October 1938)*; X:4 (October 1940); XI:2 (April 1941); XII:2 (April 1942)*; XIII:4 (October 1943)*; XIV:2 (April 1944)*; XV:1 (February 1945)*; XVI:4 (October 1946)*; XVII:2 (April 1947)*; XVIII:1 (February 1948); XIX:4 (October 1949)*; XX:2 (April 1950); XXX:1 (February 1951); XXII:4 (October 1952); XXV:4 (October 1955); XXVIII:4 (October 1958).

CURRICULUM: I:1 (January 1931); IV:2 (April 1934); VII:2 (April 1937)*; XII:3 (June 1942); XV:3 (June 1945); XVIII:3 (June 1948); XXI:3 (June 1951); XXIV:3 (June 1954); XXVI:2 (April 1956); XXVII:3 (June 1957).

EDUCATIONAL MEASUREMENT: II:3 (June 1932); II:4 (October 1932); III:1 (February 1933); V:3 (June 1935); V:5 (December 1935); VIII:3 (June 1938); VIII:5 (December 1938); XI:1 (February 1941)*; XIV:1 (February 1944)*; XVII:1 (February 1947)*; XX:1 (February 1950)*; XXIII:1 (February 1953); XXVI:1 (February 1956)*; XXIX:1 (February 1959).

EDUCATIONAL PSYCHOLOGY: I:4 (October 1931); I:5 (December 1931); II:1 (February 1932); III:4 (October 1933); IV:5 (December 1934); V:1 (February 1935); VI:3 (June 1936)*; VII:5 (December 1937)*; VIII:1 (February 1938); IX:3 (June 1939); XVIII:6 (December 1948).

EDUCATIONAL SOCIOLOGY: VI:4 (October 1936); VII:1 (February 1937); IX:4 (October 1939)*; X:1 (February 1940); XIII:1 (February 1943); XVI:1 (February 1946)*; XIX:1 (February 1949); XXII:1 (February 1952); XXIII:4 (October 1953); XXV:1 (February 1955); XXVIII:1 (February 1958).

GUIDANCE AND COUNSELING: III:3 (June 1933); VI:2 (April 1936)*; IX:2 (April 1939)*; XII:1 (February 1942)*; XV:2 (April 1945)*; XVIII:2 (April 1948*; XXI:2 (April 1951); XXIV:2 (April 1954)*; XXVII:2 (April 1957).

MENTAL AND PHYSICAL DEVELOPMENT: III:2 (April 1933); VI:1 (February 1936); IX:1 (February 1939); XI:5 (December 1941)*; XIV:5 (December 1944); XVII:5 (December 1947)*; XX:5 (December 1950); XXII:5 (December 1952); XXV:5 (December 1955); XXVIII:5 (December 1958).

LANGUAGE ARTS, FINE ARTS, NATURAL SCIENCES, AND MATHEMATICS: X:2 (April 1940); XI:4, Part 1 (October 1941); XII:4 (October 1942)*; XIII:2 (April 1943)*; XV:4 (October 1945)*; XVI:2 (April 1946); XVIII:4 (October 1948); XIX:2 (April 1949); XXI:4 (October 1951); XXII:2 (April 1952); XXV:2 (April 1955); XXVII:4 (October 1957); XXVIII:2 (April 1958).

RESEARCH METHODS: IV:1 (February 1934); IX:5 (December 1939)*; XII:5 (December 1942); XV:5 (December 1945)*; XVIII:5 (December 1948); XXI:5 (December 1951)*; XXIV:5 (December 1954)*; XXVI:3 (June 1956); XXVII:5 (December 1957).

SPECIAL PROGRAMS: VI:5 (December 1936); X:5 (December 1940); XI:3 (June 1941)*; XI:4, Part 2 (October 1941)*; XIII:5 (December 1943); XIV:3 (June 1944)*; XIV:4 (October 1944)*; XVI:5 (December 1946); XVII:3 (June 1947)*; XVII:4 (October 1947); XIX:5 (December 1949); XX:3 (June 1950)*; XX:4 (October 1950); XXIII:2 (April 1953); XXIII:3 (June 1953)*; XXIII:5 (December 1953); XXIV:1 (February 1954); XXIV:4 (October 1954); XXVI:4 (October 1956); XXVI:5 (December 1956); XXVII:1 (February 1957).

TEACHER PERSONNEL: I:2 (April 1931); IV:3 (June 1934); VII:3 (June 1937)*; X:3 (June 1940); XIII:3 (June 1943); XVI:3 (June 1946)*; XIX:3 (June 1949)*; XXII:3 (June 1952); XXV:3 (June 1955); XXVIII:3 (June 1958).

Forthcoming Issues

THE EDUCATIONAL PROGRAM: EARLY AND MIDDLE CHILDHOOD. April 1959. A. W. Foshay, *Chairman.*

ADULT EDUCATION. June 1959. Burton W. Kreitlow, *Chairman.*

HUMAN RELATIONS. October 1959. W. W. Charters, *Chairman.*

---

* Out of print.